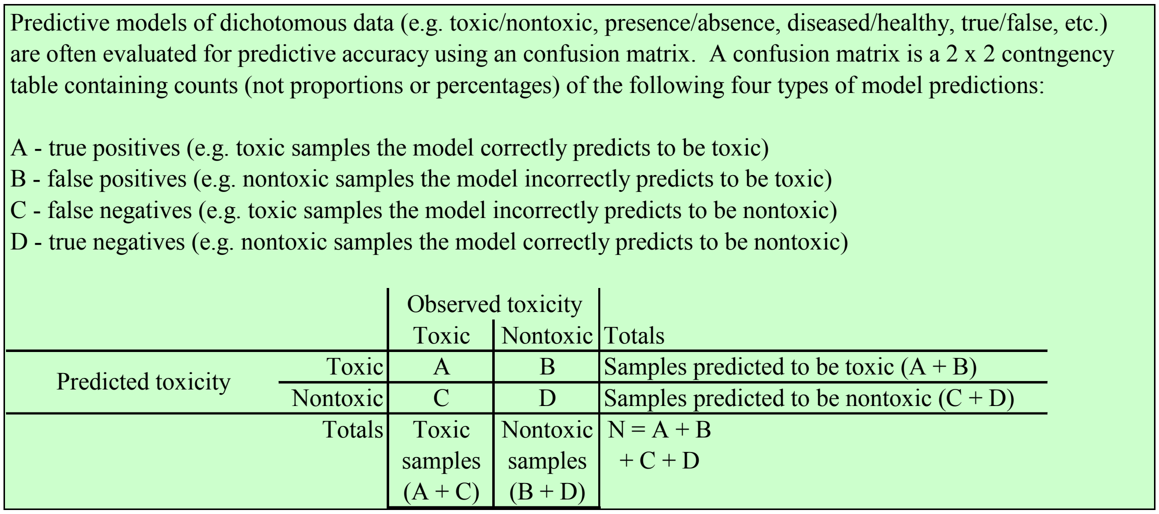
Attachment 2
Reliability Statistics for Evaluating Sediment Toxicity Predictive Models

It is surprisingly difficult to provide an adequate definition of the reliability of the sediment toxicity predictive models. Reliability in the context of the BERA is a general term indicating the level of agreement between predicted toxicity and true toxicity as represented by empirically measured sediment toxicity test results. The best test of a model’s predictions is its future performance in predicting toxicity at locations currently without empirical toxicity data. The LWG subjected several lines of evidence in the BERA to reliability analyses to determine their utility in the BERA. These include the bulk sediment quality benchmarks such as probable effect concentrations (PECs), the floating percentile model (FPM) predictions of sediment toxicity from bulk sediment chemistry data, and the logistic regression model (LRM) predictions of the probability of toxicity from bulk sediment chemistry data.

It should first be recognized, as EPA has discussed elsewhere, that elimination of a line of evidence and its findings from the conclusions of the BERA based on a perceived lack of reliability is a risk management decision. Risk management decisions within the BERA are inappropriate and unacceptable to EPA. All lines of evidence must be fully discussed and identified during risk characterization. All chemicals of concern identified by all lines of evidence in the draft BERA, including those rejected by LWG due to reliability concerns must be forwarded to the feasibility study. It is acceptable to discuss model reliability in terms of the uncertainty associated with model predictions.

Nearly all reliability estimates under discussion in the BERA are ultimately derived from categorizing model predictions against empirically measured toxicity in a contingency table. A contingency table may have any number of rows and columns of two or more. In the BERA, there are four levels of toxicity with two levels of effect (a 2 x 4 contingency table). For simplicity of discussion the examples herein are based on a 2 x 2 contingency table, with the four possible model predictions as shown in Figure 1.



Predictive models of dichotomous data (e.g. toxic/nontoxic, presence/absence, diseased/healthy, true/false, etc.) are often evaluated for predictive accuracy using an confusion matrix. A confusion matrix is a 2 x 2 contngency table containing counts (not proportions or percentages) of the following four types of model predictions:

A - true positives (e.g. toxic samples the model correctly predicts to be toxic)
B - false positives (e.g. nontoxic samples the model incorrectly predicts to be toxic)
C - false negatives (e.g. toxic samples the model incorrectly predicts to be nontoxic)
D - true negatives (e.g. nontoxic samples the model correctly predicts to be nontoxic)

		Observed toxicity		
		Toxic	Nontoxic	Totals
Predicted toxicity	Toxic	A	B	Samples predicted to be toxic (A + B)
	Nontoxic	C	D	Samples predicted to be nontoxic (C + D)
	Totals	Toxic samples (A + C)	Nontoxic samples (B + D)	N = A + B + C + D

Figure 1. Example 2 x 2 contingency table showing the four possible model predictions (true positives, false positives, false negatives, true negatives).

Literature describing reliability of predictive models such as those used in the BERA is found in many scientific disciplines. Reliability is commonly discussed in the fields of ecology, medical diagnostics, and meteorology/climatology. Thus, the literature cited in this comment is taken primarily from these scientific fields, and is applied to the reliability questions at hand in the BERA. The terminology used by statisticians in these fields is first presented in Table 1, as it differs in several areas from the terminology used by the LWG. Also, EPA has identified several additional reliability metrics not evaluated in the BERA that we believe possess advantages over the reliability metrics evaluated by LWG. All reliability statistics discussed in this comment are defined in Table 1.

Table 2 describes what question each reliability statistic answers, along with its range of possible values and some description of how to interpret each statistic. Two of the most commonly used reliability metrics in the statistical literature for evaluating predictive model accuracy are the correct classification rate (called overall reliability in the BERA) and the kappa statistic (Looney 2002), which was not used in the BERA.

Reliability Statistics Used in the BERA, Other Available Reliability Statistics

Attachment 6, pages 40-41 of the draft BERA describes the seven reliability statistics used by LWG in the BERA. They are: false negative rate; false positive rate; sensitivity; efficiency (more commonly called specificity in the literature); predicted hit reliability; predicted no-hit reliability; and overall reliability. As noted in the BERA, these statistics have been previously used with other sets of sediment toxicity data in the Pacific Northwest. These same statistics are also commonly used in other scientific fields.

As noted in Table 2, each of the above seven reliability statistics answers a different question about predictive model performance or sediment quality benchmark predictive accuracy. Each statistic provides useful information to EPA risk assessors and risk managers. As no one statistic provides all information needed by EPA to fully evaluate predictive model accuracy or sediment quality benchmark reliability, EPA concurs with LWG that multiple reliability statistics are needed to fully describe predictive model or sediment quality benchmark accuracy.

Many other reliability statistics can be calculated from contingency tables (Fielding and Bell 1997, Byrt et al. 1993, Glas et al. 2003, Tartaglione 2010). Among them are several variations of Cohen’s kappa, whose value when maximized is a commonly used method to evaluate logistic regression models, in the same way that maximizing a correlation coefficient is used to identify the best fitting linear regression model of a data set. The reliability statistics employed by LWG are applied to overall model accuracy, and are not well suited for making toxicity predictions at individual sediment sampling locations. Other reliability statistics not used by LWG in the BERA, such as positive and negative likelihood ratios, are useful in predicting the chances of toxicity at individual sampling locations with known sediment chemistry. This information will be particularly useful in the feasibility study. Finally, two additional benchmarks, bias and chance prediction rate, give the direction of bias of a sediment benchmark (i.e. benchmark either underpredicts or overpredicts toxicity), and the probability that a model makes correct predictions solely due to chance, respectively. Bias and chance values are particularly useful in the uncertainty analysis of the BERA.

Several of the reliability statistics in Table 1 are complements of each other (i.e. their sums equal 1.0). It is apparent that as one value in a complement pair increases, its complement must decrease. Other reliability statistics often provide the most useful information if evaluated in tandem with a second reliability estimate, as shown below.

Complement pairs

- Overall reliability, misclassification rate
- Prevalence, Overall diagnostic power
- Sensitivity, False negative rate
- Specificity, False positive rate

Evaluate in tandem

- Sensitivity, specificity
- Predicted hit / no-hit reliability
- Positive / negative likelihood ratio
- Kappa, prevalence adjusted bias adjusted kappa (PABAK)

Prevalence and What We Know About Sediment Toxicity in Portland Harbor

Discussed only briefly in the BERA, and not in the context of predictive model accuracy, is a term called prevalence. In the context of the BERA, prevalence is the true proportion of the stations with measured toxicity data that exhibit toxicity. Four levels of toxicity have been defined in the BERA (Levels 0, 1, 2 and 3; or no, low, moderate and severe toxicity). For feasibility study purposes, toxic is defined as the number of stations exhibiting moderate or severe toxicity (Level 2 or 3), while nontoxic stations have no or low toxicity (Level 0 or 1). As will be shown, prevalence has a major impact on the values and interpretation of reliability estimates generated by LWG in the draft BERA.

The Portland Harbor BERA has a total of 293 stations with co-occurring toxicity and sediment chemistry data. These 293 stations constitute the empirical data set used to develop the site specific versions of the logistic regression model and floating percentile model. Toxicity prevalence varies among the four toxicity endpoints (*Chironomus dilutus* survival and biomass, *Hyalella azteca* survival and biomass) as shown in Table 3.

Table 3. Number of stations exhibiting toxicity and prevalence of toxicity (in parenthesis) among the four sediment toxicity test endpoints in the BERA.

Test	Level 0 (no toxicity)	Level 1 (low)	Level 2 (moderate)	Level 3 (severe)
Chironomus survival	188 (0.642)	54 (0.184)	19 (0.065)	32 (0.109)
Chironomus biomass	201 (0.686)	37 (0.126)	12 (0.041)	43 (0.147)
Hyalella survival	253 (0.863)	19 (0.065)	2 (0.0068)	19 (0.065)
Hyalella biomass	167 (0.570)	53 (0.181)	43 (0.147)	30 (0.102)

Sediment Quality Benchmark Effects on Reliability Statistics

It seems obvious that a change in the value of a sediment quality benchmark will alter the value of reliability statistics (Figure 2). Within a given toxicity data set, as the value of a benchmark increases, the number of stations identified as false positives (i.e. the number of nontoxic stations incorrectly identified as toxic) will decrease, but the number of false negative stations (i.e. the number of toxic stations incorrectly identified as not causing toxicity) will increase (Figure 2C). Conversely, if the value of the benchmark is lowered, the number of false positive values will increase, at the expense of a decrease in the number of false negative values (Figure 2B). The extent of the increase or decrease is dependent not only on the selected value of the sediment quality benchmark, but also on the amount of overlap between sediment chemical concentrations associated with toxicity and the presumably lower sediment concentrations that do not elicit toxicity (Figure 2D).

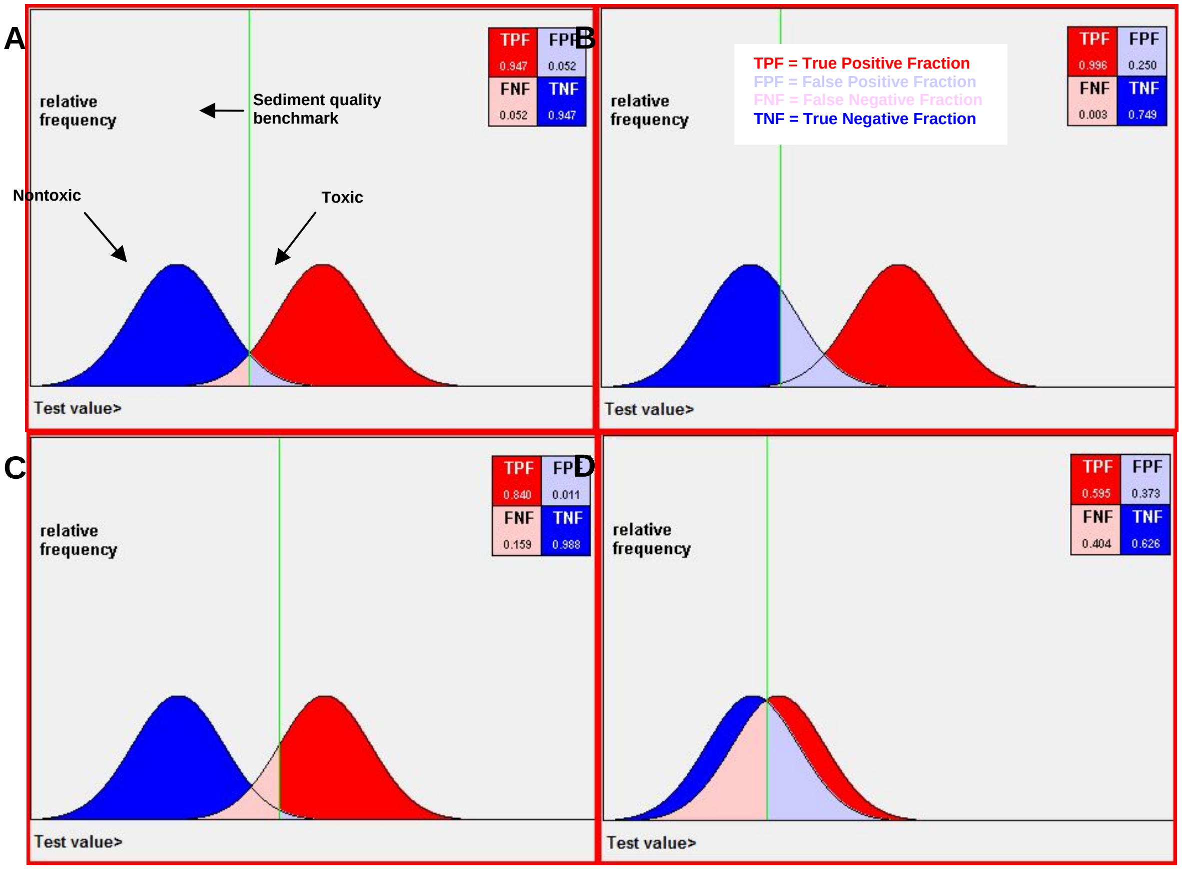


Figure 2. Sediment quality benchmark threshold effects on some reliability statistics. A. Benchmark with good ability to separate toxic from nontoxic stations, as shown by low proportion of both false positives and false negatives; B. Low benchmark resulting in few false negatives, but a high proportion of false positives; C. High benchmark resulting in few false positives, but a high proportion of false negatives; D. Site with little separation between sediment chemistry associated with toxic and nontoxic stations, meaning even a good model or benchmark has little ability to discriminate between toxic and nontoxic.

The changes in the proportion of true positives, true negatives, false positives and false negatives caused solely by changes in the value of a sediment quality benchmark (Figure 2) illustrate the information needed to identify the most reliable and accurate models or sediment quality benchmarks in the BERA. EPA believes the most reliable and accurate predictive models and benchmarks maximize the agreement between predicted and measured toxicity, while minimizing the number of incorrect predictions of toxicity. Thus, the most reliable model is unlikely to be one that meets a predefined value or range of one or more reliability statistics.

The BERA is not a competition between multiple approaches of estimating sediment toxicity to benthic biota, with the winner being the most reliable. Each of the three primary lines of evidence evaluating sediment chemistry (bulk sediment chemistry benchmarks such as PECs, logistic regression models and floating percentile models) provide different information to EPA risk assessors and risk managers. Sediment quality benchmarks provide information about sediment chemical concentrations associated with adverse effects to benthic biota at other locations in North America. Logistic regression models provide information about the probability of toxicity to benthic biota from mixtures of chemicals. The floating percentile model provides information about predicted error rates of sediment benchmarks derived from organisms exposed to mixtures of chemicals.

The different information obtained from each of these three lines of evidence is a primary reason EPA required their inclusion in the problem formulation for the BERA. At a site where sediment remediation is likely to be the primary means of reducing ecological risks to benthic biota, the risk characterization conclusions and their uncertainties from all three lines of evidence must be reported in the final BERA, because part of EPA's risk assessment and risk management determinations will be made based on concordance between these multiple lines of evidence.

Conditional Probabilities

The values of the reliability statistics calculated by the LWG, as well as the additional reliability statistics identified by EPA in Table 1 are dependent on one or more factors. One obvious factor affecting reliability statistics is the previously discussed threshold value of the sediment quality benchmark used to divide toxic concentrations from nontoxic concentrations. Changes in the cell counts of a contingency table (Figure 1) will affect the calculated values of any reliability statistic for which the value of one or more cell counts changes if the cell(s) whose count changes is in the formula to calculate the reliability statistic (Table 1).

Many, but not all of the reliability statistics in Table 1 depend on the prevalence of toxicity in the datasets used to develop the predictive models. This can be demonstrated by rewriting the formulas used to calculate reliability statistics (Table 1) in terms of prevalence. As one example, the commonly used overall reliability statistic can be written in terms of prevalence, sensitivity and specificity (Fielding and Bell 1997), in addition to the simpler calculation shown in Table 1 (sum of correctly predicted toxic stations plus correctly predicted nontoxic stations divided by the total number of stations with measured toxicity data).

Floating percentile model runs begin with the user defining an allowable false negative rate. By changing false negative rates in different FPM runs, the cell counts within a contingency table must change to account for the change in the predefined false negative rate, and thus the values of the calculated reliability statistics for each FPM run also change.

The dependence of many draft BERA reliability statistics on other events or factors places the LWG's reliability statistics within the realm of conditional probabilities. Conditional probability is the probability of an event given (or contingent upon) another event that has already occurred, e.g. the probability of a toxic effect Y on a species given prior exposure of the species to chemical X.

In the BERA, the fact that the reliability statistics used by the LWG are conditional probabilities is important only to the extent that the user of reliability statistics needs to be aware of the following. Changes in the sediment quality benchmark value and/or the prevalence of toxicity will result in changes in the calculated values of the various reliability statistics in the BERA. Prevalence effects on the calculated values of reliability statistics, if not accounted for or at least acknowledged, will affect the meaning, interpretation and utility of the reliability statistics.

However, conditional probability statistics become very important within the Portland Harbor feasibility study. This is because conditional probabilities can help answer risk management questions about sediment quality benchmarks in the feasibility study. Specifically, conditional probability statistics can be used to provide guidance in answering questions such as the following: "if a chemical sediment quality benchmark is exceeded at a sampling station without measured toxicity data, what is the probability that station would be toxic to benthic biota and requires remediation?"

EPA expects description of model and benchmark uncertainties to be the primary use of reliability statistics in the BERA. Identification of predictive models and sediment quality benchmarks that maximize the agreement between predicted and measured toxicity (i.e. simultaneously minimize both false positives and false negatives) is also a valid use of reliability statistics in the BERA. Use of reliability statistics to eliminate evaluation of lines of evidence, predictive models or sediment quality benchmarks to quantify risks is not an acceptable use of reliability statistics in the BERA.

Predictive Model Calibration and Validation

Two additional terms require definition before predictive model reliability can be fully evaluated. They are model calibration and model validation. Calibration is the process by which model parameters, predictive variables and/or the model structure itself is altered to produce better agreement between model predictions and an empirical dataset. Statisticians term the empirical dataset used to calibrate a model the "gold standard" or "training set". This is because it represents the real world situation the model is trying to replicate. The ideal goal for model calibration is to develop a predictive model with 100% accuracy (i.e. all cases are correctly classified, with no false positives or false negatives).

Validation occurs after a model is calibrated. Model validation compares model predictions from a calibrated model to measured results from a dataset not used to develop the model. The EPA recommended approach for evaluating benthic risk in the BERA (MacDonald and Landrum 2008) suggested one of two methods be used to validate the LRM and the FPM: validation with a subset of the data excluded from the original development of the models; or to split the existing

293 stations with toxicity data into two subsets, with one being used to calibrate, the other used to validate the calibrated models.

Without any effort in the BERA to validate either the FPM or LRM predictions, the BERA makes an implicit assumption that the overall reliability of sediment quality benchmarks derived from both the FPM and the LRM to predict toxicity at stations without empirical toxicity data will be equal to the overall reliability of the models as calculated from the 293 stations used to derive the models.

As summarized by Olden et al. (2002), statisticians have long known using the same data to both calibrate and validate a model leads to an overstated and biased estimate of model reliability. This is because the calibrated model has been optimized to evaluate the unique characteristics and variability in the training data set (the gold standard), and therefore loses predictive ability beyond the calibration dataset. A closely related issue, and one germane to Portland Harbor, is the situation where the training data set and the data set on which the model is to be used to predict toxicity (i.e. Portland Harbor stations without measured toxicity) are similar. In this situation, the apparent predictive accuracy of a model will also be overestimated. This overestimation of accuracy reflects the ability of the predictive model to reproduce the input to the model, rather than the model's ability to interpolate and extrapolate toxicity in a second, independent data set.

As both the site specific floating percentile and logistic regression models used all 293 stations with co-occurring toxicity and sediment chemistry data during model development in the draft BERA, both of these models are calibrated but not validated with Portland Harbor data. Generic sediment quality benchmarks, such as probable effect concentrations, have been developed with non-Portland Harbor data, and are being applied to Portland Harbor in an effort to identify unacceptable ecological risks. The generic sediment quality benchmarks are the only sediment benchmarks in the draft BERA whose predictive accuracy can be validated using empirical Portland Harbor data.

Prevalence Affects and Can Bias Conditional Probability Reliability Statistics

Several investigators have noted that if, in a model calibration dataset, the prevalence of the two endpoints is equal (i.e. 50% of samples are toxic, 50% are nontoxic), any of the reliability measures in Table 1 do a reasonably good job of describing model predictive accuracy (Fielding and Bell 1997, Manel et al. 2001, Freeman and Moisen 2008). But as shown in Table 3, the model calibration datasets for Portland Harbor do not have equal prevalence of toxic and nontoxic samples. Toxicity prevalence as designated for the feasibility study for the four sediment toxicity endpoints is shown in Table 4.

Toxicity test	**Count of toxic stations**	**Prevalence of Level 2 plus Level 3 toxicity**
Chironomus survival	49 / 293	0.167 (16.7%)
Chironomus biomass	55 / 293	0.188 (18.8%)
Hyalella survival	21 / 293	0.072 (7.2%)
Hyalella biomass	73 / 293	0.249 (24.9%)

Table 4. Prevalence of toxic effects of sediment contaminants as applied to the Portland Harbor feasibility study. Toxic is defined as the sum of the number of stations with Level 2 and Level 3 effects which are statistically significantly elevated above control.

Several investigators (Olden et al. 2002, Fielding and Bell 1997) have found that as prevalence increasingly departs from a 1:1 ratio in a model calibration dataset, the effects of prevalence on the values of many of the predictors of model reliability shown in Table 1 becomes increasingly large. In some situations, the predictions of model accuracy can also become statistically biased. The result of the prevalence and bias effects on the values of measures used to evaluate model predictive accuracy is that measures not adjusted for or which take into account these effects become increasingly poor predictors of model reliability.

Prevalence effects on three of the reliability statistics used by the LWG (overall reliability, predicted hit reliability and predicted no-hit reliability) can be directly shown by rewriting the formulas for calculating these three statistics (Table 1) in terms of prevalence (predicted hit reliability = positive predictive power or PPP; predicted no-hit reliability = negative predictive power or NPP.

Overall reliability = [(prevalence) x (sensitivity)] + [(1 – prevalence) x (specificity)]

$$\text{PPP} = \frac{\text{prevalence x sensitivity}}{\text{prevalence x sensitivity + (1 - specificity) x (1 - prevalence)}}$$

$$\text{NPP} = \frac{\text{specificity x (1 - prevalence)}}{\text{(1 - sensitivity) x prevalence + specificity x (1 - prevalence)}}$$

The remaining four LWG employed reliability statistics consist of two complement pairs: sensitivity and false negative rate, and specificity and false positive rate. Any effect of prevalence on one of the complement pair statistics will also affect the other statistic in that complement pair.

Effects of prevalence on sensitivity, specificity and their complements, false negative and false positive rates, is more difficult to demonstrate. Some investigators believe that sensitivity and specificity are not affected by prevalence (e.g. Allouche et al. 2006). Other investigators (e.g. Bruner et al. 2002c) believe they have shown sensitivity and specificity are affected by prevalence. However, prevalence effects on the values of sensitivity and specificity have been mathematically demonstrated by Choi (1997) for predictive tests.

Choi (1997) defines a predictive test in epidemiology as the situation where a positive test indicates the presence of a risk factor that causes a diseased state. Choi (1997, page 82) goes on to provide an example of a predictive test as when a chemical exposure causes a gold standard positive (a subsequent disease). If one replaces the word disease with toxicity, Choi's definition of a predictive test exactly describes the use of sediment predictive toxicity models in Portland Harbor: a chemical exposure which, if it exceeds a sediment quality benchmark, would be expected to elicit some level of toxicity in benthic biota.

At least two literature reviews of multiple predictive models, one of species presence-absence models in ecology (Manel et al. 2001), the second of medical diagnostic tests (Whiting et al. 2004) have also found that prevalence affects both sensitivity and specificity. Prevalence would also affect the complements of sensitivity and specificity, the false negative and false positive rates. Both reviews found that in the real world, prevalence affects the sensitivity and specificity of models. Both reviews found an increase in sensitivity as prevalence increased (i.e. models do a better job of predicting toxicity as the prevalence of toxicity increases in the data set). When prevalence decreases, Manel et al. 2001 found that true negatives were more effectively predicted as prevalence decreased (i.e. specificity increased as prevalence decreased), while Whiting et al. 2004 found mixed results on prevalence effects on specificity.

Finally, Choi (1997) provided a possible explanation of why different researchers have come to different conclusions regarding whether or not sensitivity and specificity (or their complements false negative and false positive rates) are affected by prevalence. Choi (1997) proposed that the differences may be due to confounding by one or more of the underlying risk factors.

The practical implications of prevalence effects on the values calculated from the reliability statistics evaluated by the LWG are as follows. Consider the definitions of three of the reliability statistics evaluated by the LWG:

- Correct classification rate (overall reliability) is the proportion of stations whose results were correctly predicted (either as toxic or nontoxic)

- Positive predictive power (PPP or predicted hit reliability) is the proportion of stations eliciting toxicity that are correctly predicted.

- Negative predictive power (NPP or predicted no-hit reliability) is the proportion of stations not eliciting toxicity that are correctly predicted.

These proportions by themselves are of only limited utility, however. This is because the predictive value of these three reliability statistics directly depends on the prevalence of toxicity at the 293 stations tested; which may well differ from the prevalence at the remaining Portland Harbor stations without measured toxicity data.

The rarer toxicity is in a data set (i.e. the lower the prevalence) the more sure we can be that a negative test indicates no toxicity, but the less sure we can be that a positive result really indicates the presence of toxicity. This is the reason that, in low prevalence data sets such as Portland Harbor's, low sediment toxicity benchmarks (e.g. threshold effect concentrations or TECs) indicating the absence of toxicity if not exceeded appear to be more reliable than higher sediment quality benchmarks (e.g. probable effect concentrations or PECs), which are designed to indicate the presence of toxicity if exceeded.

If the prevalence of toxicity is very low, the positive predictive power (predicted hit reliability) cannot be close to its maximum value of 1, even if both the sensitivity and specificity of a predictive model or sediment quality benchmark are high. The implication of this for low prevalence data sets such as those from Portland Harbor is that it is inevitable that a number of locations predicted to be toxic will be false positives.

The conclusion that prevalence can affect the value of all seven of the reliability statistics used by the LWG in the BERA can therefore be directly demonstrated for three of the reliability statistics, and is supported by both theoretical mathematical and applied observational studies for the remaining four reliability statistics.

Use of reliability statistics not affected by prevalence (or ones which can be adjusted to account for prevalence) to evaluate predictive model or sediment benchmark accuracy is highly desirable in both the BERA and feasibility study. This is because of the primary intended use of both predictive toxicity models and sediment quality benchmarks, which is to estimate toxicity or risk to benthic biota at locations without measured toxicity data. Selecting a predictive toxicity model or sediment quality benchmarks with statistics not dependent on prevalence means that a model or benchmarks derived from the 293 stations with co-occurring toxicity and sediment chemistry data can easily be transferred to other parts of Portland Harbor with a different prevalence of toxicity in the population. Such a transfer should result in no change in the accuracy of model or benchmark predictions than those derived from the stations with co-occurring toxicity and sediment chemistry data. This ability eliminates the need for assuming that toxicity prevalence at locations in Portland Harbor without empirical toxicity data is the same as it is in the 293 stations with measured toxicity data.

Another way of saying this is that selection of a predictive model or sediment quality benchmark based on conclusions from statistics whose values are affected by prevalence is less desirable, because the reliability of the model or benchmarks will change with any change in prevalence of toxicity. Such models or benchmarks may not be generally applicable harborwide, because at the very least, the reliability of the model or benchmarks will change with changes in prevalence. This is an undesirable situation if we wish to have confidence in predictions of toxicity from a model or sediment quality benchmarks throughout the entire harbor, not just at the 293 locations with co-occurring sediment toxicity and chemistry data.

Bias

Bias, termed systematic error in the epidemiology literature, favors particular results. A predictive model is biased if it systematically overpredicts or underpredicts an outcome. Within the BERA, a predictive model would be biased if it consistently tends to either overpredict or underpredict toxicity. In other words, a model is biased if the sediment quality benchmark concentrations derived from the model are either too low, resulting in an overprediction of the number of stations exhibiting toxicity, or are too high, in which case the number of stations exhibiting toxicity is underpredicted.

The magnitude of bias within a particular model is difficult to calculate, as it is dependent on several factors specific to a given data set and model run, including the selected threshold (i.e. selected sediment quality benchmark), sensitivity, specificity and prevalence (Freeman and Moisen 2008, Allouche et al. 2006, Gambino 1997). In particular, the reliability of models derived from low prevalence data sets is particularly sensitive to the selection of the threshold (Freeman and Moisen 2008). In practical terms, this means that a small change in a sediment quality benchmark for a data set with low prevalence of toxicity can result in a large change in the apparent accuracy of the model.

However, a relatively simple bias estimator can be calculated from a contingency table (Table 1) that gives the direction of bias, if any, of a given model. The bias estimator given in Table 1 answers the following question: How similar are the frequencies of predicted toxicity from a model and the frequency of observed empirical toxicity in the dataset used to develop the model (the gold standard)? It indicates whether the predictive model has a tendency to underpredict or overpredict toxicity, with the tendency to under- or over predict increasing as the bias estimator increasingly departs from unity. The bias estimator has a range from 0 to ∞, and is interpreted as follows:

Bias < 1: Toxicity underpredicted (sediment quality benchmark too high)
Bias = 1: No bias
Bias > 1: Toxicity overpredicted (sediment quality benchmark too low)

There are several reasons for the effects of prevalence on the values of model predictive accuracy measures. A primary one is that none of the reliability measures evaluated by LWG utilize all of the information regarding model predictive accuracy that is available in a confusion matrix (Fielding and Bell 1997). This loss of information can skew reliability measures. Another type of information loss occurs in the situation where toxicity prevalence is low, the case for all four Portland Harbor sediment toxicity data sets. The problem is that a severely unbalanced data set does not contain sufficient information to allow one to distinguish between an excellent model and a more mediocre predictive model (Hripcsak and Heitjan 2002). Note that this is not an issue of sample size, but instead is an issue of the relative proportions of toxic vs. nontoxic samples in the gold standard used to derive the predictive model.

The effect of the latter information loss (i.e. low prevalence) can be demonstrated with the commonly used overall reliability measure (correct classification rate). Overall reliability works well as a predictor of model accuracy when the four contingency table cell counts (true positives [A], false positives [B], false negatives [C] and true negatives [D]) are comparable to each other, but tends towards unity (i.e. 100% overall reliability) when D >> (A + B + C), irregardless of actual model performance (Delitala 2005). This last situation occurs for all four Portland Harbor sediment toxicity tests, as stations with no (Level 0) or low (Level 1) toxicity drastically outnumber stations exhibiting moderate (Level 2) or severe (Level 3) toxicity (Table 3).

Unfortunately, all of the reliability measures evaluated by LWG in the BERA are subject to this prevalence effect and the potential for statistical bias, and are not as useful as reliability measures as they would first appear. The prevalence and bias effects will affect the values of the predictive accuracy estimation statistics of all models and sediment quality benchmarks used in the BERA. Thus, our comments on reliability statistics should not be construed as a criticism of any particular predictive model or sediment benchmark in the BERA.

Example of Prevalence Effects on Reliability Estimates

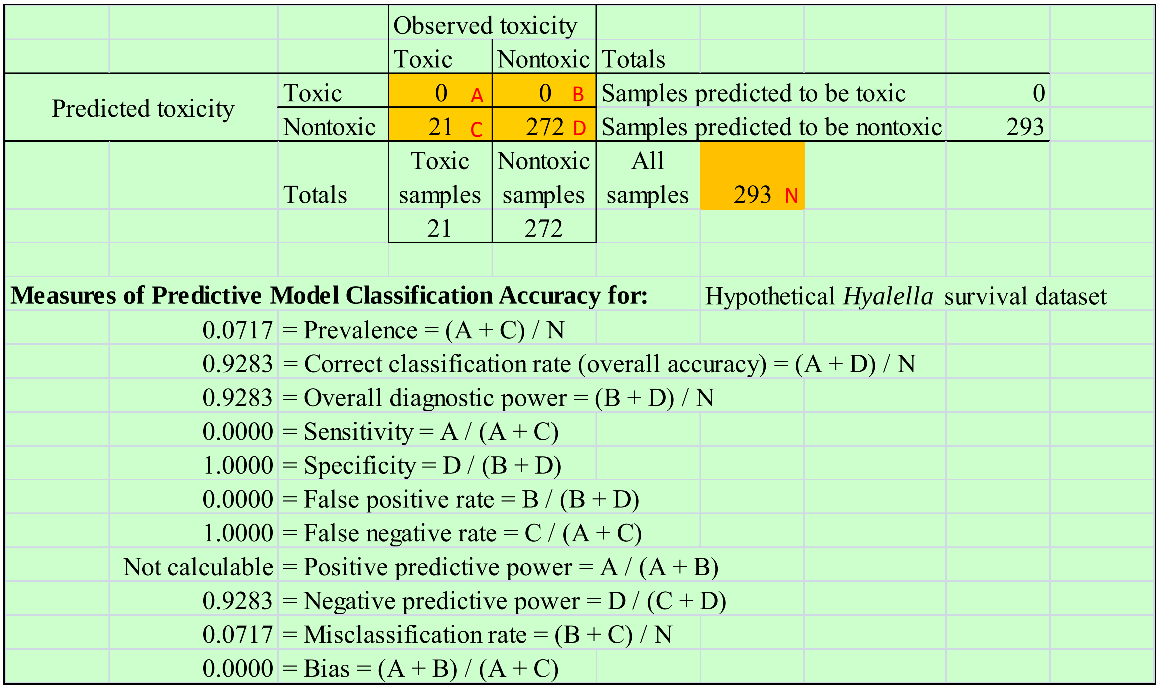
Because of the prevalence effect on the values of reliability statistics, several statisticians (Sim and Wright 2005, Lantz and Nebendahl 1996) have recommended that prevalence be reported in addition to the values of reliability statistics. An extreme example of the prevalence effect on reliability statistics can be demonstrated with the *Hyalella azteca* survival results, where only 21 out of 293 samples exhibit moderate or severe toxicity (toxicity prevalence of 7.2% as toxicity is defined in the feasibility study).

The draft BERA proposed reliability goals for sediment quality benchmarks are given on pages 40 – 41 of Attachment 6 of the draft BERA. The LWG did not discuss these reliability goals with EPA prior to submission of the draft BERA, nor were they agreed to by EPA prior to submission of the draft BERA. The reliability goals were as follows:

- Correct classification rate (overall reliability) > 80%
- Negative predictive power (predicted no-hit reliability) > 90%
- False positive rate < 20%
- False negative rate < 20%

For the *Hyalella* survival data from Portland Harbor, it would be possible to obtain a overall reliability[1] of 92.8% simply by defining a sediment quality benchmark higher than any chemical concentration at a nontoxic station, therefore correctly classifying all nontoxic stations as nontoxic, but also so high that it would incorrectly classify all empirically measured toxic stations as nontoxic (Figure 3).

[1] Overall reliability's dependence on prevalence is easily demonstrated by rewriting the overall reliability equation (Table 1) as its equivalent: [(Prevalence) x (Sensitivity)] + [(1 – Prevalence) x (Specificity)].



		Observed toxicity			
		Toxic	Nontoxic	Totals	
Predicted toxicity	Toxic	0 A	0 B	Samples predicted to be toxic	0
	Nontoxic	21 C	272 D	Samples predicted to be nontoxic	293
	Totals	Toxic samples 21	Nontoxic samples 272	All samples	293 N

Measures of Predictive Model Classification Accuracy for: Hypothetical *Hyalella* survival dataset

0.0717	= Prevalence = (A + C) / N
0.9283	= Correct classification rate (overall accuracy) = (A + D) / N
0.9283	= Overall diagnostic power = (B + D) / N
0.0000	= Sensitivity = A / (A + C)
1.0000	= Specificity = D / (B + D)
0.0000	= False positive rate = B / (B + D)
1.0000	= False negative rate = C / (A + C)
Not calculable	= Positive predictive power = A / (A + B)
0.9283	= Negative predictive power = D / (C + D)
0.0717	= Misclassification rate = (B + C) / N
0.0000	= Bias = (A + B) / (A + C)

Figure 3. Hypothetical *Hyalella azteca* survival sediment benchmark for Portland Harbor derived by setting the benchmark concentration so high as to incorrectly classify all toxic samples as nontoxic, to demonstrate the effect of prevalence on reliability statistics.

Under the hypothetical situation illustrated in Figure 3, the reliability statistics used by the LWG in the draft BERA would have been calculated to be:

Correct classification rate = [0 + 272] / 293 = 0.928 (92.8% correct classification)

No-hit reliability = 272 / [21 + 272] = 0.928 (92.8% no-hit reliability)

False positive rate = 0 / [0 + 272] = 0.0 (0% false positive rate)

Under this hypothetical situation, the correct classification rate (overall reliability), predicted no-hit reliability and the false positive rate all meet the reliability goals of the draft BERA. Only the false negative rate would not meet the reliability goals of the draft BERA, as shown below.

False negative rate = 21 / [0 + 21] = 1.0 (100% false negative rate)

It can also be observed that the bias (Table 1) for the hypothetical example shown in Figure 3 is the maximum possible bias that can be calculated for the situation where a sediment quality benchmark or predictive model underpredicts toxicity (i.e. sediment quality benchmark too high). Figure 3 is an obvious example of the situation where the combination of low prevalence and a sediment benchmark that is too high results in a situation where many reliability statistics indicate acceptable model performance, while the model itself has no utility in identifying stations exhibiting toxicity.

A predictive model that sets sediment quality benchmark so high as to incorrectly classify all toxic stations as nontoxic would clearly have no utility in identifying toxic stations in either the BERA or the feasibility study. The opposite situation, where a predictive model sets sediment quality benchmarks so low as to incorrectly classify all nontoxic stations as toxic, is unlikely to occur under the low prevalence found in the Portland Harbor sediment toxicity datasets.

Sediment quality benchmarks for the *Chironomus* survival and biomass toxicity tests can also be defined to meet the draft BERA reliability goals for overall reliability, predicted no-hit reliability and false positive rate merely by defining a sediment quality benchmark so high that it incorrectly classifies nearly all toxic stations as nontoxic. This situation occurs solely because of the relatively low prevalence of toxicity in the 293 stations with co-occurring empirical sediment toxicity and sediment chemistry data. Only the *Hyalella* biomass test cannot be made to meet the draft BERA reliability goals merely by raising sediment quality benchmarks so high that most or all toxic stations are incorrectly classified as nontoxic. This is because the toxicity prevalence observed in the *Hyalella* biomass test (24.9%) is higher than the reliability goal for the false positive rate (<20%).

The ability of FPM sediment quality benchmarks derived from the three Portland Harbor sediment toxicity tests with prevalence less than 20% to meet many of the draft BERA reliability goals merely by setting a sediment quality benchmark so high that it incorrectly classifies most if not all toxic stations as nontoxic is of particular concern to EPA. This is because of the basic concept behind the FPM, which is that adjustment of individual chemical concentrations in the FPM is unidirectional in an upward direction (as described on page 135 of the draft BERA), which can only result in higher sediment quality benchmarks. Specifically, the FPM starts with a defined percentile of a data set that provides a low, predefined false negative rate, and then adjusts individual chemical concentrations upward until false positive rates are minimized, while retaining the predefined false negative rate.

LWG's Reliability Analyses Are Not True Measures of Model Predictive Accuracy

The floating percentile model is defined to meet certain reliability goals that are not risk based. Indeed, the floating percentile model cannot even be run without an *a priori* definition of a false negative rate: a management decision. Therefore, the floating percentile model is arguably best described as a risk management tool, not a risk assessment tool. Several reviewers of methods to evaluate predictive model reliability point out that reliability measures based on models meeting user specified requirements, or which intentionally account for sources of and costs associated with erroneous predictions fall into the realm of management decision methodologies (Freeman and Moisen 2008, Hale and Heltshe 2008, Liu et al. 2005, Fielding and Bell 1997).

This last point is one of the major conclusions of our review of the reliability analyses in the draft BERA: ***the LWG's application of reliability measures in the draft BERA is not a true measure of model or sediment quality benchmark reliability, because one or more reliability measures (e.g. false negative rate) have been subjectively set at predefined values to meet risk management goals.***

Reliability measures that require compliance with user specified requirements are not true measures of model predictive accuracy (Freeman and Moisen 2008). Instead, they are subjective measures of model predictive accuracy. This is because a specific value for one or more attribute (e.g. overall reliability, predicted no-hit reliability, false positive or false negative error rates) are predetermined. While forcing a model to fit within one or more predetermined attributes can be appropriate to meet management goals within a feasibility study, such a process usually results in a model or benchmark that does not represent the maximum possible agreement between the gold standard data set used to calibrate a model and the predictions of toxic and nontoxic for stations without empirical toxicity data required in the BERA.

Within the BERA, EPA expects the LWG to use objective approaches to determine predictive model accuracy and reliability. By objective approaches, EPA means that predictive models must be calibrated in such a manner that the sediment quality benchmarks derived from the models are chosen to maximize the agreement between observed and model predicted toxicity for the 293 Portland Harbor stations for which co-occurring sediment toxicity and sediment chemistry data are available.

To evaluate predictive model or sediment quality benchmark reliability, EPA believes that the most useful reliability measures for both the BERA and the feasibility study are metrics that utilize all available information from a contingency table, not just a subset of the available information. Such measures do not suffer from the information loss that is inherent to the reliability measures used in the draft BERA. Reliability measures also need to take into account the relatively low prevalence of toxicity in the four sets of sediment toxicity data from Portland Harbor.

Reliability Measures That Are Not Affected by Prevalence

One solution to the effect of prevalence on reliability estimates is to base model reliability evaluations on accuracy measures that can either be adjusted for prevalence, or whose values are not dependent on prevalence in the calibration dataset. Such statistics also have the useful property of assessing the extent to which models correctly predict toxicity at rates that are better than chance predictions of accuracy. Finally, to avoid information loss from not using all available information in a contingency table, a reliability statistic would need to be calculated using information in all cells in a contingency table. A number of such statistics with these properties exist, but were not discussed or evaluated in the draft BERA.

EPA believes that the additional statistics that can be derived from a contingency table (Table 1) and which have the above properties are in many respects superior measures of model and sediment quality benchmark predictive accuracy compared to the reliability statistics evaluated by LWG in the BERA. These statistics are presented and discussed in the next several sections. We wish to reiterate that all of the reliability statistics referred to in this comment, both those used by the LWG and the additional statistics presented by EPA provide useful information to risk assessors and risk managers. Each reliability statistic also answers a different question (Table 2). While all reliability statistics provide useful information, the advantages and shortcomings of each reliability statistic with respect to evaluating model and sediment quality benchmark predictive accuracy must be recognized.

Agreement Between Predicted and Measured Toxicity Expected by Chance

Any predictive model or sediment quality benchmark, no matter how good or poor, will make some number of correct predictions due solely to chance. The most accurate and most useful predictive models and sediment quality benchmarks in the BERA are those that maximize the number of correct predictions over and above the number of correct predictions expected solely by chance.

In order to describe predictive accuracy in terms of improvement over correct predictions obtained by chance, the proportion of all the correct predictions due solely to chance must be known. Fortunately, the cell counts in a contingency table can be used to calculate the expected agreement between predicted and measured toxicity obtained by chance (Table 1). Chance is the level of agreement expected between predicted and measured toxicity if a predictive model or sediment quality benchmark randomly classified stations as toxic or nontoxic.

A commonly asked question is if modeled or sediment quality benchmark predictions of toxicity are better than those that can be obtained by chance. A naïve answer to this question involves flipping a coin to decide whether a station is toxic or nontoxic, which at first glance would appear to result in a 50% overall reliability rate. This naïve approach does not yield the right answer to the question regarding the number of correct predictions of toxicity due to chance, because the number of correct chance predictions is related to prevalence (Olden et al. 2002). The exact calculation of the chance agreement probability (Table 1) is among the more complex calculations derived from a contingency table (Fignre 1), but can be approximated by the simple formula[2] below.

Overall reliability due solely to chance probability ≈ 0.5 + (0.5 – prevalence)

For the low prevalence toxicity data sets from Portland Harbor, chance agreement of correct predictions of toxicity will actually be greater than 50%. Thus, the low prevalence provides upper limits or constraints on the improvement over chance agreement any Portland Harbor predictive model or sediment quality benchmark can provide.

Odds Ratio

One simply calculated statistic with the desirable properties of independence from prevalence, provides information on improvement of predictions over chance predictions, and calculated from all information in a contingency table is the odds ratio (Table 1). The odds ratio is commonly used in epidemiology, where it is used to express the strength of association between exposure and disease (Glas et al. 2003). The odds ratio is increasingly used in ecology (Manel et al. 2001). Odds ratio can be defined as the ratio of the odds of toxicity in samples predicted to be toxic relative to the odds of toxicity in samples predicted to be nontoxic. The odds ratio appears to be unaffected by prevalence (Glas et al. 2003, Fielding and Bell 1997), and also provides an

[2] The approximate formula for the overall reliability due to chance probability assumes that under the null hypothesis that a predictive model performs no better than random assignment of toxic and nontoxic predictions, the number of correctly classified cases approximates a binomial distribution (Olden et al. 2002).

indicator of the improvement of a model or benchmark in predicting toxicity above chance predictions (Manel et al. 2001). Unfortunately, the odds ratio cannot be calculated if one or more of the cells in a contingency table contains a count of zero.

Odds can be defined as the ratio of the denominator of a probability or proportion to the numerator of the probability or proportion. For example, if the probability of a horse winning a race is 50% (0.5 or ½), the odds of the horse winning are 2:1.

The odds ratio is interpreted as follows. As an example, a particular model or sediment quality benchmark results in the odds ratio being calculated as 6.2. An odds ratio of 6.2 indicates that the odds for a prediction being correct that a station elicits toxicity is over 6 times greater than the odds for a prediction that a station elicits toxicity is incorrect. The odds ratio is a measure of relative risk, meaning that it evaluates the general concept of comparing risks of toxicity at stations exposed to higher contaminant concentrations to toxicity risks at stations exposed to lower contaminant concentrations.

Cohen's kappa

A statistic believed by some to be minimally affected by prevalence effects is Cohen's kappa (usually just called kappa). Kappa is commonly used to evaluate logistic regression models (Liu et al. 2005, Looney 2002). Unlike linear regression, where the best model fit to data can be identified by maximizing either a correlation coefficient (r) or coefficient of determination (r^2), logistic regression model output has no direct analog to r or r^2. Instead, the best fitting of a series of logistic regression models is often identified as the model which maximizes the value of kappa.

Kappa maximization is commonly used in the fields of ecology and medical diagnostics as a measure of logistic regression model accuracy, and has been used to evaluate the accuracy of logistic regression based models of sediment toxicity (Bay et al. 2008). More generally, kappa maximization is commonly used in the evaluation of contingency tables. Mathematically, Feinstein and Cicchetti (1990) have demonstrated that the value of kappa can be affected by prevalence. In practice, however, some (e.g. Feinstein and Cicchetti 1990), but not all statisticians (e.g. Manel et al. 2001) have been able to demonstrate that the value of kappa is affected by prevalence that departs from 50%. Given the differences in the literature regarding the effect, or lack thereof, of prevalence on the value of kappa, the magnitude of prevalence effects on kappa may be model and application specific. Given the widespread use of kappa in the scientific literature, particularly in evaluating logistic regression, and the fact that it uses all information available in a contingency table, and thus does not suffer from information loss, EPA believes kappa is a worthwhile statistic to evaluate during predictive model and sediment quality benchmark accuracy assessment in the uncertainty analysis of the BERA. A pictoral representation of what values of kappa represent is presented in Figure 4.

A potential problem with relying solely on kappa as a measure of model reliability was first described by Feinstein and Cicchetti (1990). They observed that in some situations the phenomenon that a model would result in a high correct classification rate (high overall reliability), indicative of a good model, but the value of the kappa statistic would be low, indicative of a poorly performing model. The resolution of this apparent paradox proposed by Cicchetti and Feinstein (1990) was to not rely on a single measure of accuracy, but instead evaluate multiple measures of accuracy to obtain a more complete picture of accuracy. As Table 2 shows that each statistic that can be derived from a contingency table answers a different question, the recommendation of Cicchetti and Feinstein (1990) to look at multiple metrics when evaluating reliability appears to be a sound recommendation.

Hanssen-Kuipers Discriminant

Allouche et al. (2006) have demonstrated mathematically that kappa is a special case of a more generally applicable statistic called the Hanssen-Kuipers discriminant (Hanssen and Kuipers 1965) that is not affected by changes in prevalence. Specifically, Allouche et al. (2006) showed kappa to be the special case of Hanssen-Kuipers when prevalence is 50%. The more generally applicable Hanssen-Kuipers discriminant can be used to evaluate model reliability for any prevalence of an adverse effect.

Hanssen-Kuipers is commonly used in meteorology and climatology to evaluate the predictive accuracy of weather and climate models, and is interpreted in the same way that values of kappa are interpreted. The Hanssen-Kuipers discriminant is called the Youden's J statistic in some of the older medical diagnostic literature, and has recently been called the true skill statistic (TSS) by ecological modelers. All three terms refer to the same statistic. Figure 4 also is a conceptual representation of what different values of the Hanssen-Kuipers discriminant mean.

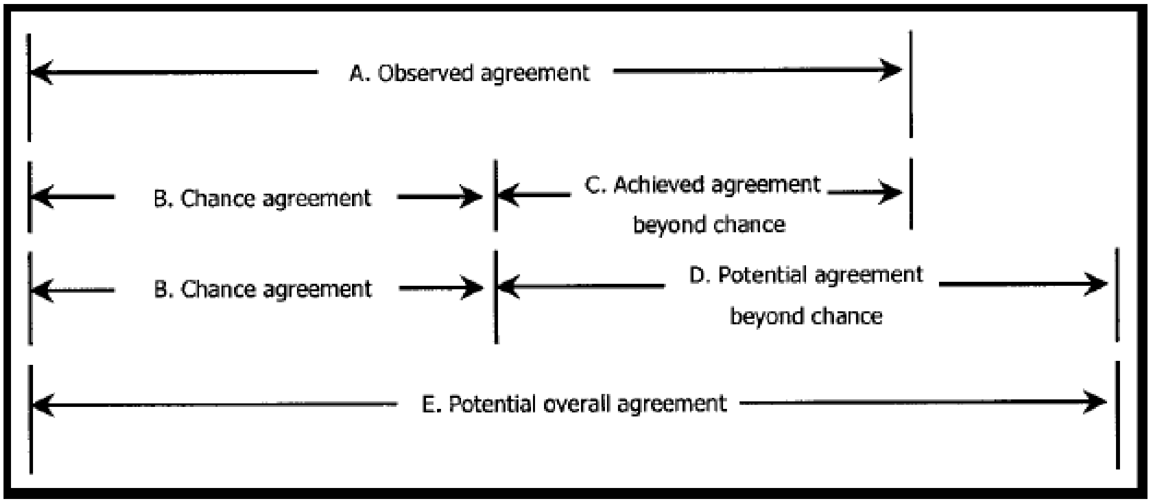

A. Observed agreement
B. Chance agreement
C. Achieved agreement beyond chance
B. Chance agreement
D. Potential agreement beyond chance
E. Potential overall agreement

Figure 4. Schematic relationship between kappa or the Hanssen-Kuipers discriminant to overall and chance agreement of model predictive accuracy. Modified from Sim and Wright (2005).

The calculated value of a Hanssen-Kuipers discriminant is not believed to be affected by the prevalence within a dataset (Woodcock 1976, Allouche et al. 2006). As such, it is a reliability statistic whose value is unaffected by prevalence, thus making it a useful reliability measure under conditions where other reliability measures are skewed or biased by prevalence effects.

Prevalence Adjusted Bias Adjusted Kappa (PABAK)

Because both prevalence and bias may play a part in determining the value of kappa, adjustments have been proposed to account for possible bias and prevalence effects on kappa. Kappa can be adjusted for prevalence by computing the average of contingency table cells A and D (Figure 1) and substituting this average for the actual counts in those cells. Similarly, a bias adjustment is performed by substituting the mean of contingency table cells B and C (Figure 1) for those actual cell counts. The kappa coefficient that results is referred to by its developers (Byrt et al. 1993) as the prevalence adjusted bias adjusted kappa (PABAK).

Hoehler (2000) criticized the use of PABAK because he believes that the effects of bias and prevalence on the magnitude of kappa are themselves informative and should not be adjusted for and thereby disregarded. Thus, Hoehler (2000) believes PABAK could generate a kappa value unrelated to the conditions for which the original predictive toxicity models were developed and would be applied. Therefore, the PABAK coefficient on its own may be uninformative, providing little or no information, because it relates to a hypothetical situation in which no prevalence or bias effects are present. However, if PABAK is presented in addition to, rather than in place of, the obtained value of kappa, Hoehler (2000) considered its use appropriate. Specifically, Hoehler (2000) believes that as part of a suite of reliability statistics including both kappa and PABAK, PABAK would give an indication of the likely effects of prevalence and bias alongside the kappa value derived from the specific measurement context studied. Hoehler, therefore, is another statistician who recommends that model predictive accuracy be evaluated using multiple statistical measures.

Normalized Mutual Information (NMI)

Unlike the other reliability measures in the draft BERA and this comment, the normalized mutual information (NMI) statistic originated in the field of information theory (Forbes 1995). In simplest terms, the NMI describes how much of the total available information in a contingency table is lost by a predictive model or sediment quality benchmark. A perfect NMI score of one means that a model or benchmark captures all of the information in a data set. Formally, it is the difference between the overall information contained in the contingency table and that in the predictions, divided by the information contained in the observed toxic vs. nontoxic data (for Portland Harbor, the 293 stations with measured sediment toxicity data), all taken from one (Forbes 1995).

Mathematically, the NMI is a measure of information based on Shannon’s entropy. In information theory, entropy is a measure of the uncertainty associated with a random variable. Shannon entropy is a measure of the average information content one is missing when one does not know the true value of the random variable. One application of Shannon entropy familiar to most biologists is the Shannon-Wiener species diversity index, with which the NMI shares many mathematical properties. The Shannon-Wiener species diversity index is also based on Shannon entropy, and summarizes community structure based on two properties of the community: species richness and species evenness. If one considers a sediment sampling location within Portland Harbor, we begin with no knowledge of whether or not a site is toxic to benthic biota. If we have a predictive model of toxicity or a sediment quality benchmark and measured sediment chemistry, we have some information about toxicity. But the information we have is uncertain, and we cannot be fully certain about toxicity at the site unless we go out and perform a toxicity test at the site. This uncertainty is what the NMI attempts to describe.

The NMI is not affected by prevalence in the data set (Manel et al. 2001, Fielding and Bell 1997). The NMI shares one disadvantage with the odds ratio. If one or more of the cells of a contingency table contain a zero value, the NMI statistic cannot be calculated.

Interpretation of Kappa, Hanssen-Kuipers Discriminant and PABAK Values

Landis and Koch (1977) have proposed the following as standards for strength of agreement for the kappa coefficient in the epidemiology literature:

$\kappa \leq 0$ = poor;
κ between .01-.20 = slight;
κ between .21-.40 = fair,
κ between .41.-60 = moderate;
κ between .61-.80 = substantial; and
κ between .81-1 = almost perfect

The choice of such an interpretive framework is, however, arbitrary. The effects of prevalence and bias on kappa must be considered when judging its magnitude. The same Landis and Koch (1977) interpretive framework can also be used to evaluate the values of the Hanssen-Kuipers discriminant and PABAK calculated from the predictive toxicity models.

PABAK is an example of a reliability statistic whose value is adjusted for prevalence effects, as opposed to the Hanssen-Kuipers discriminant, which is a reliability statistic whose value is unaffected by prevalence. EPA believes that reliability statistics either not affected by the prevalence of toxicity, or which can be adjusted to account for prevalence effects (e.g. kappa, the Hanssen-Kuipers discriminant, PABAK, odds ratio) are all likely better descriptors of toxicity model predictive accuracy than are the reliability statistics proposed by LWG in the draft BERA, all of which can be affected by the low prevalence of toxicity in the Portland Harbor sediment toxicity datasets.

Methods of Selecting Accurate Sediment Quality Benchmarks

The statistics discussed to this point can all be used to interpret the reliability and predictive accuracy of both predictive models and the sediment quality benchmarks derived from the models. Although the effect of benchmark values on the reliability statistics have already been discussed (Figure 3), there has been no discussion to this point on how to select sediment quality benchmarks that maximize predictive accuracy.

For the BERA, the most accurate sediment quality benchmarks are those that simultaneously minimize both false positive and false negative predictions of toxicity. This is a different goal than a screening level ecological risk assessment (SLERA), where the goal is to conservatively identify all chemicals potentially posing unacceptable risk, at the cost of a potentially elevated number of nontoxic stations incorrectly classified as toxic (i.e. an elevated false positive rate).

A number of methods of benchmark selection have been discussed at some length in the statistical literature. Although there are numerous studies that have evaluated the statistical properties of the various reliability statistics discussed in this comment, relatively few studies have compared the possible methods of deriving thresholds to identify the threshold selection methods that result in the highest predictive accuracy. Two studies that have compared the predictive accuracy of different threshold selection methods in ecology are those of Freeman and Moisen (2008) and Liu et al. (2005).

Toxicity data is expressed on a continuous scale, e.g., mortality can take any value between 0 – 100%. The feasibility study is concerned with summarizing this range of toxicity values into two groups: toxic or nontoxic. In statistical terms, the feasibility study dichotomizes the data into two groups. Dichotomizing continuous data into two groups results in some loss of information, and is a source of uncertainty. The BERA, which breaks the toxicity data into four groups, still requires the selection of thresholds to transform the continuous toxicity data into four groups: Level 0, 1, 2 and 3, or no, low, moderate and severe toxicity.

The threshold selected to divide toxic from nontoxic samples has, as discussed earlier, a large effect on the values of reliability statistics, including how prevalence of toxicity is defined. Within the BERA, EPA requires objective approaches to evaluating both predictive model and sediment quality benchmark reliability. Objective threshold selection methods are those that maximize agreement between observed and predicted distributions of toxicity (Liu et al. 2005).

Sediment Quality Benchmark Derivation in the Draft BERA

The sediment quality benchmark derivation approach used by the LWG follows directly from their approach to evaluating toxicity predictive model reliability. As applied to the floating percentile model, the LWG started by defining an allowable false negative rate, and then increased individual chemical concentrations upward until false positive rates were minimized, while retaining the predefined false negative rate. This approach is termed the required sensitivity method of optimizing threshold values by Freeman and Moisen (2008), one of 11 threshold optimization approaches they reviewed.

The reason for the name required sensitivity method becomes clear when it is remembered that the false negative rate is the complement of sensitivity (i.e. the sum of the false negative rate and sensitivity always equals 1). In addition to an *a priori* definition of false negative rate (e.g. 0.20), the LWG approach also results in an *a priori* definition of sensitivity (i.e. 0.80 given the *a priori* definition of a false negative rate of 0.20).

Once the false negative rate and its complement sensitivity are defined, the floating percentile model increases individual chemical concentrations upwards until false positive rates are minimized, while maintaining the predefined false negative rate and sensitivity. By minimizing false positive rates, the floating percentile model also maximizes specificity (efficiency in the BERA), the complement of the false positive rate. Specificity is the true negative rate, the proportion of truly nontoxic samples correctly predicted by a model or sediment quality benchmark.

By design, the floating percentile model therefore maximizes the number of nontoxic samples correctly predicted to be nontoxic. As the FPM increases sediment concentrations, the specificity (true negative rate) increases, while concurrently the number of false positives decreases. A consequence of the floating percentile model maximizing specificity is that the number of false negatives (toxic samples incorrectly classified as nontoxic) cannot be indefinitely maintained as a constant, but instead has to begin to increase as sediment concentrations in the FPM are increased.

In a conservative risk assessment, a goal is to minimize the number of false negatives (i.e. number of truly toxic locations incorrectly classified as nontoxic), because the risk assessor wants to ensure all locations and chemicals that pose potentially unacceptable risks are identified. At some point as sediment concentrations are increased in the FPM, the number of toxic samples incorrectly classified as nontoxic will begin to increase, which is not an acceptable situation in a risk assessment.

For use in a risk assessment, a floating percentile model approach that is the opposite of what the current FPM does may be more appropriate. In such an approach, which Freeman and Moisen (2008) term the required specificity method, the model would start by defining an *a priori* false positive rate (the complement of specificity, which is why Freeman and Moisen term this threshold derivation method the required specificity method), then ***lower*** sediment chemical concentrations until the sensitivity of the model (i.e. the true positive rate) is maximized while maintaining the predefined false positive rate. Such an approach would maximize the number of toxic stations correctly predicted as toxic, a goal of a conservative risk assessment.

As currently constructed, the floating percentile model cannot maximize sensitivity, the number of toxic stations correctly classified as toxic. At best, the FPM can define a sensitivity and its complement the false negative rate. In any event, previous reviews of reliability analyses and threshold optimization approaches (Fielding and Bell 1997, Liu et al. 2005, Freeman and Moisen 2008) all concluded that reliability statistics and evaluation criteria based on a model or benchmark meeting user specified requirements are not objective measures of model or benchmark predictive accuracy, and thus are not comparable to objective predictive accuracy measures that do not require *a priori* specification of one or more values of reliability statistics. While EPA believes these subjective approaches are not appropriate measures of reliability or predictive accuracy in the BERA, they can and do have utility in the feasibility study and remedy selection phases of the Portland Harbor project. A brief discussion of how subjective reliability measures and statistics can be used to inform management decisions in the feasibility study is provided at the end of this comment.

Objective Methods for Selecting Sediment Quality Benchmark Thresholds

Both Feeeman and Moisen (2008) and Liu et al. (2005) identify multiple objective methods for selecting thresholds from dichotomized data. Unfortunately, not all of the approaches they identified are based on using the information in a contingency table to directly derive threshold values. For the purposes of this comment, EPA will limit the discussion of objective threshold selection methods to those that can be derived from a contingency table.

Several approaches of objective threshold selection are available that are based on maximizing the value of one or more of the reliability statistics described in Table 1. Among them are maximizing the overall reliability, maximizing the sum of sensitivity plus specificity, and a related approach of selecting a threshold at the point where sensitivity equals specificity. All of these methods are among the methods evaluated by both Freeman and Moisen (2008) and Liu et al. (2005), who both concluded that despite its inherent and seemingly common sense appeal, maximization of overall reliability (i.e. maximizing the sum of correctly predicted toxic and correctly predicted nontoxic samples) is not a particularly reliable approach for identifying thresholds or evaluating predictive models. The reason overall reliability maximization is not considered a particularly reliable method of defining thresholds between toxic and nontoxic concentrations derives from the adverse effect of prevalence on the values and interpretation of the overall reliability statistic.

Liu et al. (2005) and Freeman and Moisen (2008) came to somewhat different conclusions regarding the utility of maximizing the sum of sensitivity and specificity, or setting sensitivity equal to specificity. Liu et al. (2005) believe that these two approaches are among the better approaches for objectively selecting thresholds of dichotomized data. Freeman and Moisen (2008) did not rate maximizing the sum of sensitivity and specificity or setting sensitivity equal to specificity as among the better threshold optimization methods they evaluated, but they were not among the worst, either.

Kappa Maximization

As discussed previously, kappa maximization is commonly used in both medical diagnostics (Hripcsak and Heitjan 2002) and ecology (Allouche et al. 2006) to evaluate contingency table data and predictive model accuracy. It is widely used to evaluate logistic regression models, where it is one of the standard outputs of statistical software that can be used to evaluate model accuracy. Its absence from the reliability analyses used by the LWG in the draft BERA in evaluating logistic regression model reliability was the original statistical issue EPA had with the reliability analyses that led to this comment.

When applied to logistic regression, the objective use of kappa maximization to evaluate model accuracy and sediment quality benchmark selection involves selecting an optimum probability threshold based on the benchmark dividing nontoxic from toxic that maximizes kappa. This is determined by evaluating kappa values at successive probability increments across the entire probability range from 0.0 to 1.0.

For generic sediment quality benchmarks such as PECs and benchmarks derived from the floating percentile model, it is not possible to change the benchmark in order to find the contingency table resulting in the maximum number of true positives and true negatives. Instead, the benchmark is compared to the observed sediment chemistry at each of the 293 Portland Harbor stations with measured toxicity data, and the numbers of correctly and incorrectly predicted toxic and nontoxic stations are entered into the appropriate contingency table cells. Once this has been accomplished, kappa can be calculated and interpreted using the framework proposed by Landis and Koch (1977), which was presented earlier in this comment.

Freeman and Moisen (2008) found that kappa maximization had the lowest bias of all 11 of the threshold optimization methods they evaluated. This feature of kappa maximization would reduce uncertainties associated with predictive models and sediment quality benchmarks in the BERA.

Although not explicitly evaluated by either Freeman and Moisen (2008) or Liu et al. (2005), maximization of the Hanssen-Kuipers discriminant, PABAK and the NMI conceptually would also be objective methods of describing predictive model and sediment quality benchmark reliability.

Receiver Operating Characteristic (ROC) Curves

A statistical procedure commonly used in medical diagnostics to develop interpretive guidelines for test procedures is the receiver operating characteristic[3] (ROC) curve. ROC curves (Figure 5a-d) are obtained by plotting all sensitivity values (true positive fraction) on the y-axis against their equivalent (1 - specificity) values (false positive fraction) plotted on the x-axis for all available thresholds or benchmarks. The ROC curve thus generated is plotted as a curve in what is termed a unit square. The curve starts in the lower left hand corner of the unit square, rises rapidly towards the upper left hand corner of the unit square, then flattens out before ending at the upper right hand corner of the unit square. ROC curves have been previously used by Shine et al. (2003) to develop sediment quality benchmarks for metals.

A model that has perfect ability to separate toxic from nontoxic samples would plot as a vertical line starting at the lower left hand corner of the plot, going to the upper left hand corner of the plot, then become a horizontal line running from the upper left hand corner to the upper right hand corner of the plot. A model with no ability to discriminate between toxic and nontoxic would appear as a diagonal line running from the lower left hand corner to the upper right hand corner of the unit square. Models with intermediate discriminatory ability appear as arcs or curves. The ROC curve itself is comprised of threshold or benchmark values, with each point on the curve corresponding to a specific true positive-false positive pair of values.

If the ROC plot is a smooth curve, any tangent to the ROC curve identifies a particular sensitivity/specificity pair. The point of the ROC curve closest to the upper left hand corner of the plot is defined as the point where the tangent to the curve has a slope of 1.0. The point on an ROC curve where the tangent to the curve equals one represents the threshold or benchmark that does the best job of separating toxic from nontoxic samples. Many medical diagnostic benchmarks, such as the definition blood glucose levels >110 mg/dL as indicative of elevated blood glucose levels in individuals who should undergo definitive testing for diabetes (Somannavar et al. 2009) have been derived through the use of thresholds from an ROC curve.

[3] The unusual name of this statistic comes from its development by the British during World War II, when it was used to evaluate the ability of radar receiver operators to correctly separate friendly from enemy aircraft.

If the ROC plot is a stepped curve instead of the smooth curves shown in Figure 5, the equivalent sensitivity/specificity pair is found by moving a line, with slope m, from the top left corner of the ROC plot to the ROC curve. The threshold that results in the shortest possible line between the ROC curve and the upper left hand corner of the ROC unit square, termed the minimum ROC distance approach for threshold determination by Freeman and Moisen (2008) is found by the minimizing the following quantity:

$$\boldsymbol{(1 - \textit{Sensitivity})^2 + (\textit{Specificity} - 1)^2}$$

The above equation can be used to approximate the threshold of an ROC curve at the point where the tangent to the curve has a slope of 1.0.

In addition to providing an objective method for identifying thresholds or benchmarks, the ROC methodology also provides a way to measure model predictive accuracy. The area under the curve (AUC) of the ROC plot relates relative proportions of correctly classified (true positives) and incorrectly classified (false positives) in contingency table cells over all possible threshold values. This makes the area under a ROC curve a threshold–independent measure of model predictive accuracy (Pearce and Ferrier 2000).

The AUC ranges between 0.5 for models with no discrimination ability between toxic and nontoxic to 1.0 for models with perfect discrimination ability between toxic and nontoxic locations. A subjective guide for interpreting AUC values is that proposed by Swets (1988):

AUC = 0.90–1.00: excellent ability to discriminate between toxic and nontoxic
AUC = 0.80–0.90: good ability to discriminate between toxic and nontoxic
AUC = 0.70–0.80: fair ability to discriminate between toxic and nontoxic
AUC = 0.60–0.70: poor ability to discriminate between toxic and nontoxic
AUC = 0.50–0.60: failure to discriminate between toxic and nontoxic

AUC values of less than 0.5 indicate that the model tends to predict toxicity at nontoxic sites, indicating that the model parameters should be reversed.

Although ROC curves can be generated with spreadsheets, most ROC analyses are performed with statistical software designed to perform the analysis. Many of the larger statistical software packages such as Systat now contain ROC modules. Several smaller packages that specialize in ROC analyses, such as MedCalc, are also available.

Likelihood Ratios and Predictive Accuracy for Sites without Measured Toxicity Data

Consider the most common situation regarding sediment sampling locations in Portland Harbor, which is:

1. we know the concentration of chemicals in sediment, and
2. for chemicals with sediment quality benchmarks, we know if the sediment concentrations exceed their respective sediment quality benchmarks, but
3. we do not have any measured sediment toxicity data from the location, thus

4. we do not know whether the location actually elicits toxicity in benthic biota

In this common situation, we have knowledge of sediment chemistry analyses, but the ecological significance of the test results is uncertain. Consequently, the problem becomes deciding whether any given sediment chemistry analysis represents a true or false positive for toxicity if the sediment quality benchmark is exceeded, or whether the sediment chemistry analysis represents a true or false negative for toxicity if the sediment quality benchmark is not exceeded.

In the BERA, EPA makes the assumption that any sediment chemical concentration that equals or exceeds its sediment quality benchmark (i.e. the hazard quotient ≥ 1) poses some level of unacceptable risk to benthic biota. In the feasibility study, where the economic costs of remediating locations with acceptable levels of risk are substantial, and the environmental costs of not remediating areas posing unacceptable risks are also substantial, risk managers want to know how much confidence to place in a sediment quality benchmark that predicts a given location to be toxic or nontoxic to benthic biota.

For a station whose sediment chemistry exceeds one or more sediment quality benchmarks, and therefore is predicted to be toxic, the question can be addressed by calculating the ratio of true positives to false positives. The true positive rate of a model or benchmark is its sensitivity (Table 2). The true negative rate of a model or benchmark is its specificity, which means that the false positive rate is the complement of the specificity, or the quantity (1 – specificity). The ratio of the true positive rate to the false positive rate is sensitivity / (1 – specificity). This ratio is termed a likelihood ratio.

Likelihood ratios are actually odds, and are interpreted as follows. If a predictive model or sediment quality benchmark results in a contingency table yielding a likelihood ratio of 10, the value of 10 is interpreted to mean that the odds are 10:1 that a prediction of toxicity represents a station that would elicit toxicity if measured toxicity data became available from the station. In the statistical literature, the commonly seen term ":1" is usually not seen or reported, but is implied when describing a likelihood ratio.

Two types of likelihood ratios can be calculated from each contingency table (Table 1): a positive likelihood ratio (LR^+) and a negative likelihood ratio (LR^-). Positive likelihood ratios describe the odds of a station being a true positive if a benchmark or model predicts that station to be toxic. The larger the values of a positive likelihood ratio, the higher the odds are that a station truly would elicit toxicity if a toxicity test were to be performed at that station. In other words, the larger a positive likelihood value is, the more confidence one can have in drawing a conclusion that a station really would elicit toxicity.

As is the case for interpreting kappa, PABAK or the Hanssen-Kuipers discriminant, scales for interpreting likelihood ratios are somewhat subjective. Likelihood ratios measure the power of a model or benchmark to change the pre-test into the post-test probability of toxicity being present. One interpretive framework for interpreting positive likelihood ratios is the following:

$LR^+ = 1.0$: no predictive ability
$LR^+ = 1.0 – 2.0$: rarely important change from pretest to posttest probability

LR^+ = 2.0 – 5.0: small change
LR^+ = 5.0 – 10: moderate change
LR^+ = >10: large change

Conversely, a negative likelihood ratio describes the odds of a false negative if a model or benchmark predicts a location to be nontoxic. The smaller the value of a negative likelihood ratio, the better a model or sediment quality benchmark is at ruling out toxicity if a toxicity test were to be performed at that station. A LR^- value of 0.25 indicates that the odds are 1:4 (false negatives:true negatives) that a prediction of no toxicity represents a station that would exhibit toxicity if measured toxicity data were available. One interpretive framework for interpreting negative likelihood ratios is the following:

LR^- = 1.0: no predictive ability
LR^- = 0.5 – 1.0: rarely important change from pretest to posttest probability
LR^- = 0.2 – 0.5: small change
LR^- = 0.1 – 0.2: moderate change
LR^- = 0 – 0.1: large change

Likelihood ratios are unaffected by prevalence in a data set. Likelihood ratios also have an advantage over many of the other reliability statistics discussed in this comment, as they can be applied to predictions of toxicity at individual sediment sampling stations, not just to the entire population of sediment sampling stations. Thus, they should provide useful information regarding the accuracy of model and sediment quality benchmark values in toxicity predictions at individual stations without measured toxicity data.

Likelihood ratios provide information regarding predictive accuracy in their own right, but take on their greatest importance when used to make posttest predictions regarding the ability of a predictive model or sediment quality benchmark to discriminate between toxic and nontoxic stations.

Before the utility of likelihood ratios in benthic toxicity predictive accuracy can be fully appreciated, the terms pretest and posttest probability in the interpretive frameworks for likelihood ratios need to be more fully explained. The two terms are defined as follows:

- A pretest (prior) probability is an initial probability value originally obtained before any additional information is obtained.
- A posttest (posterior) probability is a probability value that has been revised by using additional information that is later obtained.

A likelihood ratio can be used to give the posttest odds of the model or benchmark prediction being correct. In the context of the Portland Harbor BERA, the pretest probability of toxicity is simply the prevalence of toxicity in each of the four sets of empirical toxicity data measured at the 293 stations with co-occurring sediment toxicity and sediment chemistry data. Posttest probabilities of toxicity will be calculated from the sediment chemistry data collected from the remaining sediment sampling stations in Portland Harbor without measured toxicity data. The term posttest in this context simply means additional sediment chemistry data collected at stations other than the 293 stations with measured toxicity data. Posttest does not mean that the sediment chemistry data was collected after the sediment toxicity tests were performed.

When combined with the bias (systematic error) statistic, which gives the direction of error for both predictive models and the sediment quality benchmarks derived from them (either overpredicting or underpredicting toxicity), the reliability statistics not affected by prevalence or which can be adjusted to account for prevalence effects should provide an appropriate description of the uncertainty associated with the predictive models and sediment benchmarks derived from them. These uncertainties should be discussed in the appropriate uncertainty sections of the BERA.

What must not be done in the BERA is to eliminate any lines of evidence or individual sediment quality benchmarks, or any chemical hazard quotients greater than or equal to one from the BERA because of a perceived lack of reliability. It is EPA's responsibility to make the risk management decisions regarding the use of any particular predictive model or sediment quality benchmark within the remainder of the remedial investigation and feasibility study for Portland Harbor. Risk management decisions will be made and documented by EPA in the feasibility study, not the BERA or the remedial investigation report. The uncertainties associated with the predictive models, as described and quantified by reliability statistics will inform EPA's risk management decisions, but will not be the sole basis for EPA's management decisions at Portland Harbor.

EPA is unaware of any one reliability metric that is superior to all others in all situations. All reliability statistics discussed in this comment have issues that should be recognized and which can affect their interpretation. Therefore, we are in agreement with LWG that multiple reliability metrics should be calculated and evaluated as measures of the accuracy of the FPM, LRM and sediment quality benchmarks in predicting sediment toxicity to benthic invertebrates at locations in Portland Harbor without empirical sediment toxicity data.

The Bridge Between the Use of Reliability Statistics in the BERA and in the Feasibility Study

Within the feasibility study, many management decisions will be made based on the information in the BERA and the rest of the remedial investigation report. One of the most important and difficult decisions to make will be the need, if any, for remediation of locations without any empirical (measured) sediment toxicity data. The answer to the question of the area to be remediated to protect benthic biota will depend in part on the reliability of sediment quality benchmarks, either those previously published, or those derived specifically for Portland Harbor in the BERA.

Overall accuracy of either the floating percentile or logistic regression models will not answer the above question. Instead, what is needed is an answer to the following general question:

> ***What is the probability of toxicity to benthic biota if a chemical concentration at a site without measured toxicity data exceeds a sediment quality benchmark?***

The answer to this question can be calculated if three pieces of information are known:

1. The prevalence of toxicity
2. What fraction of the toxic stations are correctly predicted by a sediment quality benchmark (i.e. what is the sensitivity of the benchmark), and
3. What fraction of the nontoxic stations are correctly predicted by a sediment quality benchmark (i.e. what is the specificity of the benchmark)

The above three pieces of information need to be combined in a way that allows the prediction of toxicity at stations without measured toxicity data, which at first would appear to be a difficult task.

The solution to the above question involves the use of Bayes theorem. Bayes theorem is a way of understanding how the probability that a theory is true is affected by a new piece of evidence. The theory we wish to test at Portland Harbor is whether a site is toxic to benthic biota given that a sediment quality benchmark is exceeded.

This theorem, developed by Rev. Thomas Bayes and originally published in 1763, relates the conditional probability of occurrence of an event to the probabilities of other events that have already occurred. Several different formulations of Bayes theorem exist. For Portland Harbor, we will answer the above question of the probability of toxicity at stations without measured toxicity data by using the following formula (Equation 1), which is one definition of Bayes theorem.

Equation 1:

Posttest odds = Pretest odds x likelihood ratio

Unlike the other reliability statistics discussed to this point, application of Bayes theorem produces results called posterior probabilities, which are revised probabilities based on new information. Bayes theorem forms the basis of much of our thinking in conditional probability and making "predictions" statistically based on historical data. In the case of Portland Harbor, application of Bayes theorem will permit estimates of the likelihood that a station without measured toxicity data will elicit toxicity in either *Chironomus dilutus* or *Hyalella azteca* survival or biomass based solely on exceedance of a sediment quality benchmark.

Bayesian statisticians base statistical inference on a number of philosophical underpinnings that differ in principle from classical statistical thought. First, Bayesians believe that research results should reflect updates of past research. In other words, prior knowledge should be incorporated formally into current research to obtain the best 'posterior' or resultant knowledge. Second, Bayesians believe that much can be gained from insightful prior, subjective information as to the likelihood of certain types of events. Third, Bayesians use Bayes theorem to translate probabilistic statements into degrees of belief, instead of a classical confidence interval interpretation.

Bayesian statistics can be used to answer the following type of question, which is of interest to feasibility study managers regarding the predictive accuracy of sediment quality benchmarks:

> ***If 7.2% of Portland Harbor sampling stations cause elevated mortality in Hyalella azteca (i.e. prevalence of toxicity is 7.2%), and a PCB sediment quality benchmark has a false positive rate of 0.20 and a false negative rate of 0.20, what is the probability that a random sediment station with a PCB concentration exceeding the benchmark will actually elicit increased mortality in Hyalella?***

The answer to the above question is that the post test probability of toxicity is 0.2368, roughly a one in four chance that a station without measured toxicity data, but whose sediment PCB concentration exceeds the PCB benchmark will elicit increased mortality of *Hyalella azteca*. But how was this answer obtained?

The first piece of information needed is the pretest odds of toxicity. Normally this is a difficult piece of information to obtain. Fortunately for Portland Harbor, we have the empirical toxicity test results from 293 sediment sampling stations with co-occurring toxicity and chemistry data. The pretest probability of toxicity for each of the four sets of toxicity test data (*Chironomus* survival and biomass, *Hyalella* survival and biomass) is simply the prevalence of toxicity for each test, which for feasibility study purposes is given in Table 4.

Prevalence however is a probability, not the odds of toxicity. Probability and odds can be readily converted to each other by the following equations:

Equation 2:

Probability = Odds / (Odds + 1)

Equation 3:

Odds = Probability / (1 – Probability)

By converting the measured prevalence of toxicity into the pretest odds of toxicity (Equation 3), we can then use Equation 1 to multiply the pretest odds of toxicity by the positive likelihood ratio of toxicity (Table 1) calculated from a contingency table to obtain the posttest odds of toxicity at a given station. If desired, the posttest odds of toxicity can be back transformed to the posttest probability of toxicity at the station (Equation 2).

A number of sediment sampling locations within Portland Harbor have two or more sediment quality benchmarks that are exceeded. Bayes theorem can be expanded to calculate the posttest odds or probability of toxicity at stations with two or more sediment quality benchmarks that are exceeded. As long as each chemical whose sediment quality benchmark is exceeded at a location has an available positive likelihood ratio value, Equation 1 can be expanded to incorporate multiple likelihood ratios as shown in Equation 4.

Equation 4:

$$\textbf{\textit{Posttest odds = Pretest odds x LR}}_{chemical\ 1} \text{ x } LR_{chemical\ 2} \text{ x . . . x } LR_{chemical\ n}$$

Where:

LR = positive likelihood ratio for each chemical’s sediment quality benchmark
n = number of chemicals at station whose sediment quality benchmarks are exceeded

Pretest odds used with Equation 4 are the same as those used with Equation 1: the prevalence of toxicity for whichever of the four available sets of measured toxicity data (*Chironomus* survival or biomass, *Hyalella* survival or biomass) is under evaluation.

The posttest odds calculated by Equation 4 are limited by the important assumption that each of the individual sediment quality benchmark likelihood ratios are independent of each other. Equation 4 permits estimation of posttest odds of toxicity at any sampling location without measured toxicity data that has any number of chemicals whose concentrations exceed their respective sediment quality benchmarks. As before, odds can be back transformed into probability of toxicity if desired. A sampling station with multiple chemicals that exceed their sediment quality benchmarks will have a higher probability of eliciting toxicity in benthic biota than will a station with only one sediment quality benchmark that is exceeded.

Two Additional Issues that May Affect All Reliability Statistics

Spatial Autocorrelation

One predictive model uncertainty not discussed to this point, but one that will affect all reliability statistics for all models and sediment benchmarks is the issue of spatial autocorrelation. Spatial autocorrelation occurs when the presence, absence, or degree of a certain characteristic at one location affects the presence, absence, or degree of the same characteristic in neighboring locations. In the context of Portland Harbor, it is the tendency of a sampling station to possess characteristics, such as chemical concentrations, the mixtures of chemicals present, and physical characteristics such as grain size distribution, that are more similar to those of their nearest neighboring sampling stations, and less similar to stations further away. Spatial autocorrelation is a potential problem for all area-based studies (Fielding and Bell 1997). If sample data are spatially autocorrelated, the assumption of independence between samples is violated, leading to problems with the significance of test statistics. Spatial autocorrelation can arise when the probability of toxicity at one location is not independent of the probability of toxicity at neighboring stations. Spatial autocorrelation effects are likely to result in reliability statistics values that overpredict the reliability of all sediment toxicity models and sediment quality benchmarks used in the BERA.

In practical terms, spatial autocorrelation may provide some useful information for managers in the feasibility study. If two stations are predicted to elicit a similar level of toxicity, but one station is adjacent to an area of potential concern (AOPC) boundary, while the second station is distant from any AOPC, the proximity of the first station to an AOPC may provide qualitative support for adjusting the AOPC boundary to include the area around the station predicted to be toxic.

Scaling to Contingency Tables Larger than 2 x 2

To this point, most comments and discussions have been limited to 2 x 2 contingency tables. As noted early in this comment, the BERA is evaluating four levels of toxicity (no, low, moderate and severe) as opposed to the two levels of toxicity (toxic or nontoxic) of concern in the feasibility study.

Fortunately, to evaluate predictive accuracy of models that evaluate multiple levels of toxicity, or sediment quality benchmarks dividing no from low, low from moderate, and moderate from high toxicity in the BERA, nearly all of the statistics presented in Table 1 can be scaled upward to evaluate multiple categories. Included in the reliability statistics that can be expanded to evaluate 2 x 4 contingency tables are some of the more complex measures such as kappa (Fleiss 1981, Fleiss 1971). Computation of reliability statistics from contingency tables larger than 2 x 2 are given in several sources, including Liu et al. 2007. The primary statistics discussed in this comment that cannot be expanded to evaluate contingency tables larger than 2 x 2 are the tangent of an ROC curve and the area under the curve of the ROC curve.

One additional, simply applied method that could evaluate the 2 x 4 contingency tables from the BERA would be to dichotomize them, then evaluate as a series of 2 x 2 contingency tables. In practice, this would mean evaluating benchmarks that would be thresholds between Level 0 and Level 1 toxicity, thresholds between Level 1 and Level 2 toxicity, and thresholds between Level 2 and Level 3 toxicity. The feasibility study definition of toxicity evaluated extensively in this comment is nothing more than an evaluation of sediment quality benchmarks from a 2 x 4 contingency table dichotomized to evaluate thresholds between the sum of Level 0 plus Level 1 toxicity, and the sum of Level 2 plus Level 3 toxicity.

Example Calculations of Reliability Statistics for Generic Sediment Quality Benchmarks, Logistic Regression Derived Benchmarks and Floating Percentile Model Derived Benchmarks

Need final results of model outputs to perform. Pick as an example a chemical such as total PCB that has benchmarks from all three sources of benchmarks (generic SQBs, LRM, FPM)

Inclusion of Remedial and Environmental Costs in Reliability Statistics

This last section of our comment on reliability statistics discusses the incorporation of remedial cost data or management goals into reliability statistics and the derivation of sediment quality benchmarks. In the BERA, EPA has assumed that the effects of false positive and false negative errors in predicting sediment toxicity are equal. Risk managers may decide that it is more important to them to, for example, develop conservative remedial goals that are protective of human health and the environment, at the cost of mistakenly requiring remediation of a few locations that would not elicit toxicity if empirical toxicity data were available.

As the weighting of costs requires management decisions not within the purview of risk assessors, we do not perform any calculations of reliability or sediment quality benchmark derivation weighted to account for costs. Instead, we identify a general procedure for incorporating costs and management goals into the calculation of reliability statistics, and provide several literature citations so that managers with interest in using costs and management goals to come to their final decisions on sediment quality benchmarks can perform their detailed analyses of costs.

Several investigators (Zweig and Campbell 1993, Fielding and Bell 1997, Liu et al. 2005) discuss the use of ROC plots in making management decisions that incorporate the weighting of costs or management goals in threshold or benchmark derivation. To do so, an initial management decision must be made of the costs of false positive and false negative errors. Assigning values to these costs is complex, at least partially subjective, dependent upon the context within which the sediment quality benchmark will be used, and falls into the realm of risk management decisions. As a guideline Zweig and Campbell (1993) suggest that if false positive costs (FPC) are greater than false negative costs (FNC), or FPC > FNC, the threshold (sediment quality benchmark) should favor specificity (i.e. the selected threshold should become less conservative, and should maximize the proportion of truly nontoxic samples that are correctly predicted). Sensitivity should be favored if FNC > FPC (i.e. the selected threshold should become more conservative, and should maximize the proportion of truly toxic samples that are correctly predicted). Estimation of the ratio of false positive to false negative costs will also serve the purpose if actual dollar amounts cannot be assigned to the costs of false positives and false negatives.

Once costs of false positive and false negative errors (or their ratio) have been defined, the prevalence (P) of toxicity is combined with the cost information, allowing the calculation of a slope (Zweig and Campbell 1993).

$m = (FPC/FNC) \times ((1-P)/P)$

If the ROC plot is a smooth curve, m describes the slope of a tangent to this curve. The point at which the tangent touches the curve identifies a particular sensitivity/specificity pair. This point is the sediment quality threshold or benchmark that is weighted after costs have been taken into account.

If the ROC plot is a stepped non-parametric curve the equivalent sensitivity/specificity pair is found by moving a line, with slope m, from the top left of the ROC plot. The sensitivity/specificity pair is found where the line and the curve first touch (Zweig & Campbell 1993). Again, the point on the curve where the line touches the ROC curve is the sediment quality threshold or benchmark that is weighted to take the costs of false positive or negative errors into account.

Summary, Conclusions and Recommendations

Despite the length and inherent complexity of some of the statistical discussions in this comment, the conclusions and recommendations from this evaluation can be summarized in relatively few points.

1. No one reliability statistic is available that provides all information needed by risk assessors and risk managers to evaluate and utilize sediment quality benchmarks and predictive models of toxicity. Multiple reliability statistics are needed.

2. The LWG's application of reliability measures in the draft BERA is not a true measure of model or sediment quality benchmark reliability, because one or more reliability statistics (e.g. false negative rate) were subjectively set at predefined values to meet LWG-proposed risk management goals.

3. Within the BERA, EPA expects the LWG to use an objective approach to determine predictive model and sediment quality benchmark accuracy and reliability. By objective approach, EPA means that predictive models must be calibrated in such a manner that the sediment quality benchmarks derived from the models maximize the agreement between observed and model predicted toxicity for the 293 Portland Harbor stations for which co-occurring sediment toxicity and sediment chemistry data are available.

4. Prevalence of toxicity, defined as the proportion of stations in the four sets of measured sediment toxicity data (*Chironomus dilutus* survival and biomass, *Hyalella azteca* survival and biomass) from 293 sampling locations in Portland Harbor that actually elicit toxicity, is low, ranging between 7 – 25% of stations eliciting either Level 2 (moderate) or Level 3 (severe) toxicity, depending on which test one is discussing. The low percentage of stations eliciting toxicity is an encouraging finding of the BERA, as it means between 75 – 93% of the 293 stations either elicit Level 0 (no toxicity), or Level 1 (low) levels of toxicity.

5. The low prevalence of toxicity in the 293 stations with co-occurring sediment chemistry and sediment toxicity data used to develop the site specific floating percentile and logistic regression predictive toxicity models adversely affects the calculated values of the reliability statistics presented in the draft BERA, as well as their interpretation, and can also bias the reliability statistics.

6. The prevalence effect is a statistical problem that directly results from the low number of stations eliciting toxicity. The problem is not due to a lack of sampling data, nor is it a criticism of any particular predictive modeling approach or sediment quality benchmark.

7. All reliability statistics evaluated by the LWG in the draft BERA, as well as all additional reliability statistics recommended for use by EPA can be derived from a contingency table that tabulates the number of true positive, true negative, false positive and false negative predictions of toxicity made by any predictive model or any individual sediment chemical benchmark calibrated with or validated against the 293 stations with measured toxicity and sediment chemistry data.

8. The BERA is not a competition between multiple lines of evidence or approaches of estimating sediment toxicity to benthic biota, with the winner being the most reliable. Each of the three primary lines of evidence evaluating sediment chemistry (bulk sediment chemistry benchmarks such as PECs, logistic regression models and floating percentile models) provide different information to EPA risk assessors and risk managers, which is why each was included in the problem formulation for the BERA.

9. The risk characterization conclusions, including all hazard quotient calculation results and their uncertainties from all three lines of evidence (generic sediment quality benchmarks, floating percentile model, logistic regression model) must be reported in the final BERA, because part of EPA's risk assessment and risk management determinations will be made based on concordance between these multiple lines of evidence. Use of reliability statistics in the BERA to eliminate lines of evidence or individual sediment quality benchmarks from risk analysis and risk characterization is unacceptable to EPA.

10. EPA expects description of predictive model and sediment quality benchmark uncertainties to be the primary use of reliability statistics in the BERA. Identification of models and benchmarks that maximize the agreement between predicted and measured toxicity (i.e. simultaneously minimize both false positives and false negatives) is also a valid use of reliability statistics in the BERA.

11. EPA's recommended solution to the effect of prevalence on reliability statistics is to base model reliability evaluations primarily on reliability statistics that can either be adjusted for prevalence, or whose values are not dependent on prevalence in the calibration dataset.

12. Many of the statistics unaffected or unbiased by prevalence also have the useful property of assessing the extent to which models correctly predict toxicity at rates that are better than chance predictions of accuracy.

13. To avoid information loss from not using all available information in a contingency table, reliability statistic are available that can be calculated using information from all contingency table cells. The reliability statistics used by LWG in the draft BERA, while providing useful information, do not make use of all available information in contingency tables.

14. Statistics not utilized by the LWG in the draft BERA, but which are unaffected by prevalence or can be adjusted to account for prevalence effects, utilize all information in a contingency table, and/or which describe the improvement of model or benchmark predictions over the agreement between predicted and measured toxicity expected solely by chance include the odds ratio, Cohen's kappa, prevalence adjusted bias adjusted kappa (PABAK), the Hanssen-Kuipers discriminant, the normalized mutual information (NMI) statistic, and likelihood ratios. These statistics and the other statistics identified by EPA but not used in the draft BERA, should be used in addition to and in conjunction with the reliability statistics used by LWG to obtain a more complete and accurate picture of model and benchmark reliability.

15. Of particular use in the BERA uncertainty analysis will be the statistic called bias, which identifies whether a model or benchmark systematically over- or underestimates toxicity, as well as identifying the direction of the bias.

16. Of particular use in the feasibility study will be the statistics positive likelihood ratio and negative likelihood ratio. Through the use of Bayes theorem, likelihood ratios, when combined with the measured prevalence in each of the four sets of sediment toxicity data, can be used to estimate the odds and/or the probability that an individual sampling station without measured toxicity data, but where one or more sediment quality benchmark is exceeded will be truly toxic. Nearly all other reliability statistics in both the draft BERA and this comment only provide probabilities for the population of all sampling stations. Bayes theorem used in conjunction with likelihood ratios and prevalence give the odds or probability that an individual station predicted to elicit toxicity would in fact cause adverse effects if toxicity were actually measured at the station.

17. Within the feasibility study, it may be desirable to weight the reliability statistics to incorporate the cost of remediation, or to account for the economic and/or environmental costs of decision errors resulting from sediment quality benchmarks that are either too conservative, resulting in unnecessary remediation of some locations, or which are not conservative enough and thus not fully protective of human health and the environment. Many, but not all of the reliability statistics discussed in this comment can be so weighted, although we have not weighted any statistics for use in the BERA.

When combined with the bias (systematic error) statistic, which gives the direction of error for both predictive models and the sediment quality benchmarks derived from them (either overpredicting or underpredicting toxicity), the reliability statistics not affected by prevalence or which can be adjusted to account for prevalence effects should provide an appropriate description of the uncertainty associated with the predictive models and sediment benchmarks derived from them. These uncertainties should be discussed in the appropriate uncertainty sections of the BERA.

What must not be done in the BERA is to eliminate any lines of evidence or individual sediment quality benchmarks, or any chemical hazard quotients greater than or equal to one from the BERA because of a perceived lack of reliability. It is EPA's responsibility to make the risk management decisions regarding the use of any particular predictive model or sediment quality benchmark within the remainder of the remedial investigation and feasibility study for Portland Harbor. Risk management decisions will be made and documented by EPA in the feasibility study, not the BERA or the remedial investigation report. The uncertainties associated with the predictive models, as described and quantified by reliability statistics will inform EPA's risk management decisions, but will not be the sole basis for EPA's management decisions at Portland Harbor.

EPA is unaware of any one reliability metric that is superior to all others in all situations. All reliability statistics discussed in this comment have issues that should be recognized and which can affect their interpretation. Therefore, we are in agreement with LWG that multiple reliability metrics should be calculated and evaluated as measures of the accuracy of the FPM, LRM and sediment quality benchmarks in predicting sediment toxicity to benthic invertebrates at locations in Portland Harbor without empirical sediment toxicity data.

As noted in Table 2, each reliability statistic discussed in this comment answers a different question. The specific questions being asked of a sediment toxicity predictive model or sediment quality benchmark will to a large extent drive which reliability statistic(s) a user will consider or evaluate. Risk assessors and risk managers will often be asking different questions, and thus will choose to evaluate or place more weight or emphasis on a different set of statistics. The expanded list of reliability statistics available for use over and above the list provided in the draft BERA should provide risk assessors and risk managers the tools needed to answer the questions each will ask, or at the very least provide information that can be used to make informed risk assessment and risk management decisions.

Literature Cited

Allouche, O., A. Tsoar and R. Kadmon. 2006. Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS). J. Appl. Ecol. 43:1223-1232.

Bay, S.M., K.J. Ritter, D.E. Vidal-Dorsch and L.J. Field. 2008. Comparison of national and regional sediment quality guidelines for classifying sediment toxicity in California. p. 79-90 in S.B. Weisberg and K. Miller, eds. Southern California Coastal Water Research Project Annual Report 2008, Costa Mesa, CA.

Byrt, T., J. Bishop and J.B. Carlin. 1993. Bias, prevalence and kappa. J. Clin. Epidemiol. 46:423-429.

Choi, B.C.K. 1997. Causal modeling to estimate sensitivity and specificity of a test when prevalence changes. Epidemiology 8:80-86.

Cicchetti, D.V. and A.R. Feinstein. 1990. High agreement but low kappa: II. Resolving the paradoxes. J. Clin. Epidemiol. 43:551-558.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educ. Psychol. Meas. 20:37-46.

Delitala, A.M.S. 2005. Perception of intense precipitation events by public opinion. Natural Hazards and Earth System Sci. 5:499-503.

Feinstein, A.R. and D.V. Cicchetti. 1990. High agreement but low kappa: I. The problem of two paradoxes. J. Clin. Epidemiol. 43:543-549.

Fielding, A.H. and J.F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ. Conservat. 24:38-49.

Fleiss, J.L. 1981. Statistical Methods for Rates and Proportions. John Wiley & Sons, New York, NY.

Fleiss, J.L. 1971. Measuring nominal scale agreement among many raters. Psych. Bull. 76:378-382.

Forbes, A.D. 1995. Classification-algorithm evaluation: five performance
measures based on confusion matrices. J. Clin. Monit. 11:189-206.

Freeman, E.A. and G.G. Moisen. 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol. Model. 217:48-58.

Gambino, B. 1997. The correction for bias in prevalence estimation with screening tests. J. Gambling Studies 13:343-351.

Glas, A.S., J.G. Lijmer, M.H. Prins, G.J. Bonsel and P.M.M. Bossuyt. 2003. The diagnostic odds ratio: A single indicator of test performance. J. Clin. Epidemiol. 56:1129-11235.

Hale, S.S. and J.F. Heltshe. 2008. Signals from the benthos: Development and evaluation of a benthic index for the nearshore Gulf of Maine. Ecol. Indicators 8:338-350.

Hanssen, A. W. and W. J. A. Kuipers. 1965. On the relationship between the frequency of rain and various meteorological parameters. Koninklijk Ned. Meteor. Instit., Meded. Verhand. 81: 2–15.

Hoehler. F.K. 2000. Bias and prevalence effects on kappa viewed in terms of sensitivity and specificity. J. Clin. Epidemiol. 53:499-503.

Hripcsak, G. and D.F. Heitjan. 2002. Measuring agreement in medical informatics reliability studies. J. Biomed. Informatics 35:99-110.

Landis, J.R. and G.G. Koch. 1977. The measurement of observer agreement for categorical data. Biometrics 33:159-174.

Lantz, C.A. and E. Nebenzahl. 1996. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. J. Clin. Epidemiol. 49:431-434.

Liu, C., P. Frazier and L. Kumar. 2007. Comparative assessment of the measures of thematic classification accuracy. Remote Sensing of Environment 107:606-616.

Liu, C., P.M. Berry, T.P. Dawson and R.G. Pearson. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography 28:385-393.

Looney, S.W. 2002. Statistical Methods for Assessing Biomarkers. p. 81-109 in Looney, S.W., ed. Biostatistical Methods. Humana Press, Totowa, NJ.

MacDonald, D.D. and P.F. Landrum. 2008. An Evaluation of the Approach for Assessing Risks to the Benthic Invertebrate Community at the Portland Harbor Superfund Site. Preliminary

Draft. Prepared for U.S. Environmental Protection Agency, Portland, OR and Parametrix, Inc., Albany, OR, September 2008. 44 pp. plus Addendum.

Manel, S., H.C. Williams and S.J. Ormerod. 2001. Evaluating presence-absence models in ecology: the need to account for prevalence. J. Appl. Ecol. 38:921-931.

Olden, J.D., D.A. Jackson and P.R. Peres-Neto. 2002. Predictive models of fish species distributions: A note on proper validation and chance predictions. Trans. Amer. Fish. Soc. 131:329-336.

Pearce, J. and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Model. 133:225-245.

Somannavar, S., A. Ganesan, M. Deepa, M. Datta and V. Mohan. 2009. Random capillary blood glucose cut points for diabetes and pre-diabetes derived from community-based opportunistic screening in India. Diabetes Care 32:641-643.

Shine, J.P., C.J. Trapp and B.A. Coull. 2003. Use of receiver operating characteristic curves to evaluate sediment quality guidelines for metals. Environ. Toxicol. Chem. 22:1642-1648.

Sim, J. and C.C. Wright. 2005. The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. Physical Therapy 85:257-268.

Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. Science 240:1285-1293.

Tartaglione, N. 2010. Relationship between precipitation forecast errors and skill scores of dichotomous forecasts. Weather and Forecasting 25:355-365.

Whiting, P., A.W.S. Rutjes, J.B. Reitsma, A.S. Glas, P.M.M. Bossuyt and J. Kleijnen. 2004. Sources of variation and bias in studies of diagnostic accuracy. A systematic review. Annals of Internal Medicine 140:189-203.

Woodcock, F. 1976. The evaluation of yes/no forecasts for scientific and administrative purposes. Monthly Weather Review 104:1209-1241.

Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine.

Zweig, M.H. and G. Campbell. 1993. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin. Chem. 39:561-77.

Table 1. Reliability measures, their definitions and calculation. Formula terms correspond to their definitions in Figure 1.

Reliability Measure	**Equivalent BERA Measure or Term**	**Definition**	**Formula**
Total number of cases	Total sampling locations	Number of stations with co-occurring toxicity and chemistry data	$N = A + B + C + D$
Prevalence	No exact equivalent, sum of the number of Level 1, 2 and/or 3 hits divided by total sampling locations closest BERA equivalent	True proportion of stations with measured toxicity data exhibiting toxicity	$\frac{(A+C)}{N}$
Correct classification rate	Overall reliability	Proportion of all cases correctly predicted	$\frac{(A+D)}{N}$
False negative rate	False negative rate	Proportion of truly toxic samples predicted to be nontoxic	$\frac{C}{(A+C)}$
False positive rate	False positive rate	Proportion of truly nontoxic samples predicted to be toxic	$\frac{B}{(B+D)}$
Sensitivity (True positive rate)	Sensitivity	Proportion of truly toxic samples correctly predicted	$\frac{A}{(A+C)}$
Specificity (True negative rate)	Efficiency	Proportion of truly nontoxic samples correctly predicted	$\frac{D}{(B+D)}$
Positive predictive power	Predicted hit reliability	Probability of presence of toxicity given a model prediction of toxicity	$\frac{A}{(A+B)}$

Reliability Measure	**Equivalent BERA Measure or Term**	**Definition**	**Formula**
Negative predictive power	Predicted no-hit reliability	Probability of absence of toxicity given a model prediction of nontoxic	$\frac{D}{(C+D)}$
Bias (Systematic error)	Not used	Tendency of model or benchmark to over- or underpredict toxicity	$\frac{(A+B)}{(A+C)}$
Overall diagnostic power	Not used	True proportion of stations not exhibiting toxicity	$\frac{(B+D)}{N}$
Misclassification rate	Not used	Proportion of all cases incorrectly predicted	$\frac{(B+C)}{N}$
Chance agreement	Not used	Level of agreement expected between predicted and measured toxicity if model or benchmark randomly classified stations as toxic or nontoxic	$\left[\left(\frac{(A+B)}{N}\right)\times\left(\frac{(A+C)}{N}\right)\right]+\left[\left(\frac{(C+D)}{N}\right)\times\left(\frac{(B+D)}{N}\right)\right]$
Odds ratio	Not used	Ratio of correctly assigned cases to incorrectly assigned cases	$\frac{(A\times D)}{(B\times C)}$
Hanssen-Kuipers discriminant (= true skill statistic = Youden's J)	Not used	Extent to which model predicts toxicity at a rate higher than expected by chance – unaffected by prevalence	$\frac{((A\times D)-(B\times C))}{((A+C)\times(B+D))} = \left(\frac{A}{(A+C)}+\frac{D}{(B+D)}-1\right) =$ Sensitivity + Specificity - 1

Reliability Measure	Equivalent BERA Measure or Term	Definition	Formula
Kappa (= Cohen's kappa)[a]	Not used	Extent to which model predicts toxicity at a rate higher than expected by chance – not adjusted for prevalence	$\frac{\left[(A+D)-\left(\left((A+C)(A+B)+(B+D)(C+D)\right)/N\right)\right]}{\left[N-\left(\left((A+C)(A+B)+(B+D)(C+D)\right)/N\right)\right]}$
Prevalence adjusted bias adjusted kappa (PABAK)	Not used	Extent to which model predicts toxicity at a rate higher than expected by chance – adjusted for prevalence and bias	$\frac{(A+D)-(B+C)}{N}$
Normalized mutual information (NMI)	Not used	The difference between the overall information in a contingency table and the information available in model or benchmark predictions	$1-\frac{-A\ln A-B\ln B-C\ln C-D\ln D+(A+B)\ln(A+B)+(C+D)\ln(C+D)}{N\ln N-\left[(A+C)\ln(A+C)+(B+D)\ln(B+D)\right]}$
Positive likelihood ratio	Not used	Probability that a station is a true positive divided by the probability that a station is a false positive	$\frac{\left(\frac{A}{(A+C)}\right)}{\left(1-\left(\frac{D}{(B+D)}\right)\right)}=\frac{\text{Sensitivity}}{(1\text{ - Specificity})}$
Negative likelihood ratio	Not used	Probability that a station is a false negative divided by the probability that a station is a true negative	$\frac{\left(1-\left(\frac{A}{(A+C)}\right)\right)}{\frac{D}{(B+D)}}=\frac{(1-\text{Sensitivity})}{\text{Specificity}}$

Reliability Measure	**Equivalent BERA Measure or Term**	**Definition**	**Formula**
Pretest probability	Not used	Bayesian term for prevalence: True proportion of stations with measured toxicity data exhibiting toxicity	$\frac{(A+C)}{N}$
Pretest odds	Not used	Bayseian term for likelihood of toxicity at a station before the toxicity test results are known. In practical terms, prevalence expressed as odds of toxicity	$\frac{\left(\frac{A+C}{N}\right)}{\left(1-\left[\frac{A+C}{N}\right]\right)} = \frac{\text{Prevalence}}{1 - \text{Prevalence}}$
Posttest odds	Not used	Bayesian term for odds that a station elicits toxicity after toxicity tests or sediment chemistry analyses are performed	$\left(\frac{\left(\frac{A+C}{N}\right)}{\left(1-\left[\frac{A+C}{N}\right]\right)}\right) \times \left(\frac{\left(\frac{A}{(A+C)}\right)}{\left(1-\left(\frac{D}{(B+D)}\right)\right)}\right)$

Reliability Measure	**Equivalent BERA Measure or Term**	**Definition**	**Formula**
Posttest probability	Not used	Bayesian term for probability that a station elicits toxicity after information from measured toxicity tests or sediment chemistry analyses becomes available	$\dfrac{\left[\left(\dfrac{\left(\dfrac{A+C}{N}\right)}{\left(1-\left[\dfrac{A+C}{N}\right]\right)}\right)\times\left(\dfrac{\left(\dfrac{A}{(A+C)}\right)}{\left(1-\left(\dfrac{D}{(B+D)}\right)\right)}\right)\right]}{\left\{\left[\left(\dfrac{\left(\dfrac{A+C}{N}\right)}{\left(1-\left[\dfrac{A+C}{N}\right]\right)}\right)\times\left(\dfrac{\left(\dfrac{A}{(A+C)}\right)}{\left(1-\left(\dfrac{D}{(B+D)}\right)\right)}\right)\right]+1\right\}}$

[a]- Although kappa is more commonly used than Hanssen-Kuipers in the statistical literature, kappa has been shown by Allouche et al. (2006) to be the special case of the Hannsen-Kuipers discriminant when prevalence = 0.5 (i.e. half the samples are toxic, half are not toxic). The statistical literature is split regarding the magnitude of the effect of prevalence on the value of kappa. This division among statisticians, along with the relative ease with which a variance term can be calculated for kappa, allowing for tests of significant differences in model predictive accuracy among different models, are likely among the reasons for the widespread use of kappa.

Table 2. What question is answered by each reliability measure, what is the range of possible values for each measure, and some guidance to interpreting each reliability measure.

Reliability Measure	Question Answered	Range of Values	Perfect Score
Prevalence	What fraction of the total sampling locations with measured toxicity data were observed to elicit toxicity?	0 to 1	Not applicable
Correct classification rate (Overall reliability)	Overall, what fraction of predictions were correct?	0 to 1	1
False negative rate	What fraction of the observed toxic stations were incorrectly predicted to be nontoxic?	0 to 1	0
False positive rate	What fraction of the observed nontoxic stations were incorrectly predicted to elicit toxicity?	0 to 1	0
Sensitivity	What fraction of the observed toxic stations were correctly predicted?	0 to 1	1
Specificity (Efficiency)	What fraction of the observed nontoxic stations were correctly predicted?	0 to 1	1
Positive predictive power (Predicted hit reliability)	What is the probability that a station actually elicits toxicity if the model or benchmark predicts it to be toxic?	0 to 1	1
Negative predictive power (Predicted no-hit reliability)	What is the probability that a station does not elicit toxicity if the model or benchmark predicts it to be nontoxic?	0 to 1	1
Bias (Systematic error)	How did the predicted frequency of toxicity compare to the observed frequency of toxicity?	0 to ∞	1
Overall diagnostic power	What fraction of the total sampling locations were observed to not elicit toxicity?	0 to 1	Not applicable
Misclassification rate	Overall, what fraction of predictions were incorrect?	0 to 1	0
Chance agreement	How much of the agreement between predicted and measured toxicity is expected to be due solely to chance?	0 to 1	0
Odds ratio	What is the ratio of the odds of a toxic prediction being correct, to the odds of a toxic prediction being wrong?	0 to ∞	∞

Reliability Measure	Question Answered	Range of Values	Perfect Score
Hanssen-Kuipers discriminant (= true skill statistic = Youden's J)	How well did the predictive model or sediment quality benchmark separate the toxic stations from the nontoxic stations?	-1 to +1	+1
Kappa (= Cohen's kappa)	What is the proportion of agreement between model predictions and empirical toxicity data over and above agreement due solely to chance?	-1 to +1	+1
Prevalence adjusted bias adjusted kappa (PABAK)	How well did the predictive model or sediment quality benchmark separate the toxic stations from the nontoxic stations after adjustment for prevalence and bias effects?	-1 to +1	+1
Normalized mutual information (NMI)	How much of the information available in the measured toxicity data is included in model or sediment quality benchmark predictions of toxicity?	0 to 1	+1
Positive likelihood ratio	How much have the odds of toxicity increased if a predictive model or sediment quality benchmark predicts a station to be toxic?	1 to ∞	∞
Negative likelihood ratio	How much have the odds of toxicity decreased if a predictive model or sediment quality benchmark predicts a station to be nontoxic?	0 to 1	0
Pretest probability	Bayesian term for prevalence: What fraction of the total sampling locations with measured toxicity data were observed to elicit toxicity?	0 to 1	Not applicable
Pretest odds	Bayseian term for prevalence expressed as odds: What are the odds that any one of the 293 Portland Harbor stations with measured toxicity data will be observed to elicit toxicity?	0 to 1	Not applicable
Posttest odds	Bayesian term that answers the following: What are the odds that a random sediment station with a chemical concentration that exceeds its sediment quality benchmark will actually elicit increased toxicity?	0 to ∞	∞

Reliability Measure	**Question Answered**	**Range of Values**	**Perfect Score**
Posttest probability	Bayesian term that answers the following: What is the probability that a random sediment station with a chemical concentration that exceeds its sediment quality benchmark will actually elicit increased toxicity?	0 to 1	1

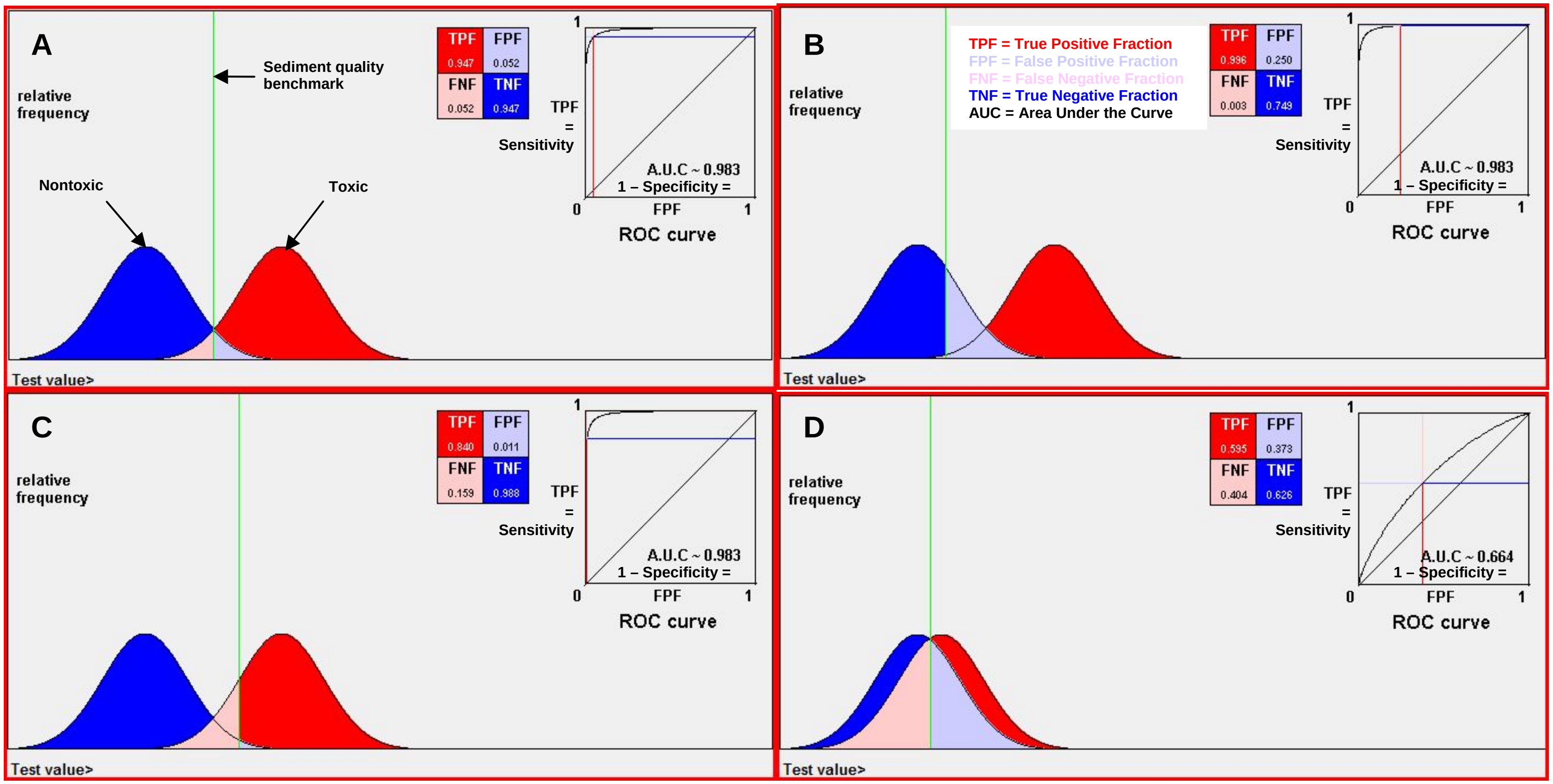


Figure 5. Use of receiver operating characteristic (ROC) curves to select sediment quality benchmarks. A. Optimal benchmark that minimizes false positives and false negatives, benchmark on the ROC curve at the point closest to the upper left hand corner of the plot; B. Low benchmark resulting in few false negatives, but a high proportion of false positives, benchmark on ROC curve nearer to upper right hand corner of the plot; C. High benchmark resulting in few false positives, but a high proportion of false negatives, benchmark on ROC curve nearer to lower left hand corner of the plot; D. Site with little separation between sediment chemistry associated with toxic and nontoxic stations, area under the curve lower than at locations with better separation between toxic and nontoxic samples. Toxicity distributions and sediment quality benchmarks same as in Figure 2.